

NAVAL POSTGRADUATE SCHOOL MONTEREY, CALIFORNIA



THESIS

**USING GENETIC ALGORITHMS TO SEARCH
LARGE, UNSTRUCTURED DATABASES: THE
SEARCH FOR DESERT STORM SYNDROME**

by

David L. Jacobson

September 1996

Thesis Advisor:

Hemant K. Bhargava

Approved for public release; distribution is unlimited.

19970123 038

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 1996	3. REPORT TYPE AND DATES COVERED Master's Thesis	
4. TITLE AND SUBTITLE USING GENETIC ALGORITHMS TO SEARCH LARGE, UNSTRUCTURED DATABASES: THE SEARCH FOR DESERT STORM SYNDROME			5. FUNDING NUMBERS	
6. AUTHOR(S) Jacobson, David L.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT Exploratory data analysis problems have recently grown in importance due to the large magnitudes of data being collected by everything from satellites to supermarket scanners. This so-called "data glut" often precludes the effective processing of information for decision-making. These problems can be seen as search problems over massive unstructured spaces. A prototypical problem of this type involves the search, by Department of Defense medical agencies, for a so-called "Desert Storm Syndrome" which involves large amounts of medical data obtained over several years following the Persian Gulf conflict. This data ranges over more than 170 attributes, making the search problem over the attribute space a hard one. We propose the use of genetic algorithms for the attribute search problem, and intertwine it with search algorithms at the detailed data level. Computational results so far strongly suggest that our system has succeeded at the given tasks, requiring relatively few resources. They also have found no indication that a single syndrome or other medical entity is responsible for wide-spread adverse health ramifications among a significant cross-section of Persian Gulf War participants in the CCEP program. There are, however, numerous correlations of exposure/demographic information and associated symptoms/diagnoses which suggest that smaller groups may share common health conditions based on shared exposure to common health risk factors.				
14. SUBJECT TERMS Desert Storm Syndrome, Genetic Algorithm, Artificial Intelligence, Medical Data Analysis			15. NUMBER OF PAGES 153	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)

Prescribed by ANSI Std. Z39-18 298-102

Approved for public release; distribution is unlimited.

**USING GENETIC ALGORITHMS TO SEARCH LARGE, UNSTRUCTURED
DATABASES: THE SEARCH FOR DESERT STORM SYNDROME**

David L. Jacobson
Lieutenant, Medical Service Corps, United States Navy
B.S., United States Naval Academy


Submitted in partial fulfillment
of the requirements for the degree of

**MASTER OF SCIENCE IN
INFORMATION TECHNOLOGY MANAGEMENT**


from the

**NAVAL POSTGRADUATE SCHOOL
September 1996**


Author:


David L. Jacobson

Approved by:


Hemant K. Bhargava, Thesis Advisor


Donald Gayer, Second Reader

 for
Reuben Harris, Chairman
Department of Systems Management

ABSTRACT

Exploratory data analysis problems have recently grown in importance due to the large magnitudes of data being collected by everything from satellites to supermarket scanners. This so-called "data glut" often precludes the effective processing of information for decision-making. These problems can be seen as search problems over massive unstructured spaces. A prototypical problem of this type involves the search, by Department of Defense medical agencies, for a so-called "Desert Storm Syndrome" which involves large amounts of medical data obtained over several years following the Persian Gulf conflict. This data ranges over more than 170 attributes, making the search problem over the attribute space a hard one. We propose the use of genetic algorithms for the attribute search problem, and intertwine it with search algorithms at the detailed data level. Computational results so far strongly suggest that our system has succeeded at the given tasks, requiring relatively few resources. They also have found no indication that a single syndrome or other medical entity is responsible for wide-spread adverse health ramifications among a significant cross-section of Persian Gulf War participants in the CCEP program. There are, however, numerous correlations of exposure/demographic information and associated symptoms/diagnoses which suggest that smaller groups may share common health conditions based on shared exposure to common health risk factors.

TABLE OF CONTENTS

I. INTRODUCTION	1
A. ANALYSIS OF LARGE DATABASES.....	1
B. PURPOSE OF THIS RESEARCH.....	2
C. SCOPE OF RESEARCH.....	3
D. REAL WORLD APPLICABILITY.....	4
E. THESIS METHODOLOGY AND ORGANIZATION.....	5
F. ACKNOWLEDGMENTS.....	7
II. COMPREHENSIVE CLINICAL EVALUATION PROGRAM	9
A. BACKGROUND AND HISTORY OF CCEP.....	9
B. CCEP RESEARCH VISION.....	11
C. DATABASE DESCRIPTION.....	12
D. WHY DOES A GENETIC ALGORITHM WORK FOR CCEP ANALYSIS?.....	13
1. Theory.....	13
2. Advantages and Disadvantages of the Genetic Algorithm Method.....	16
E. KEY CHALLENGES TO CCEP ANALYSIS BY A GENETIC ALGORITHM.....	17
1. Problem Structure.....	17
2. Database Content and Structure.....	20
3. Database Normalization.....	23
4. What is "Interesting?".....	26
III. SOLUTION CONCEPTS	37
A. RESEARCH GOALS.....	37
B. SOLUTION STRATEGY.....	39
IV. DAMI GENETIC ALGORITHM ARCHITECTURE.....	43
A. PROGRAM MODULES.....	43
1. The Genetic Algorithm Package.....	44
2. The Statistical Analysis Package.....	45
3. User Interface.....	46
B. REPORTING AND FILTERING.....	47
C. SYSTEM REQUIREMENTS.....	49
1. Hardware and Software Requirements.....	49
2. Processing Limits.....	50
V. SEARCHING THE HYPOTHESIS SPACE: DAMI IMPLEMENTATION.....	51
A. THE GENETIC ALGORITHM.....	51
B. THE STATISTICAL ANALYSIS ALGORITHM.....	53
C. TUNABLE PARAMETERS.....	54
D. PROBLEMS AND IMPROVEMENTS.....	55
1. Convergence Issues.....	55
2. Processing Speed Issues.....	57
3. Tuning the Fitness Measure, Verification, and Validation.....	58
VI. RESULTS.....	63
A. SUMMARY.....	63
B. DID THE GENETIC ALGORITHM PERFORM AS EXPECTED?.....	64
1. Analysis Speed.....	65

2. Hypothesis Quality Improvement.....	66
3. Reproducibility: Search Space Coverage.....	67
C. WHAT DID DAMI FIND?	70
1. Exposure-to-diagnosis Correlations.....	72
2. Exposure-to-symptom Correlations.....	77
D. ARE THE RESULTS USEFUL TO MEDICAL PROFESSIONALS?	80
VII. CONCLUSION.....	81
A. LESSONS LEARNED	81
B. RECOMMENDATIONS FOR FUTURE RESEARCH.....	82
APPENDIX A. CCEP DATA DICTIONARIES AND DATA COLLECTION METHODOLOGY.....	85
A. DATA DICTIONARY OF CCEP DATABASE.....	85
B. DATA COLLECTION METHODS.....	90
APPENDIX B. DATA DICTIONARY OF SELECTED DAMI FILES	93
APPENDIX C. TOP 100 HYPOTHESES DISCOVERED BY EXPOSURES-TO-DIAGNOSIS AND EXPOSURE-TO-SYMPTOM STUDIES	97
LIST OF REFERENCES.....	139
INITIAL DISTRIBUTION LIST.....	141

This thesis is dedicated to:

*My G_d who has given me the small talent I have to contribute,
My wife, Laurie, who stands beside me and sacrifices her career goals
so I can serve my country and pursue mine
My "littles," Zachary and Erin, who constantly remind me of where reality lies
and what is important in life
And to all veterans of the United States Armed Forces
who have served their country faithfully and suffer in silence...*

I. INTRODUCTION

A. ANALYSIS OF LARGE DATABASES

Twenty years ago, computers were relatively scarce and applied to limited, highly specialized applications. At that time, there were rarely enough computerized data to make them an integral part of any organization's decision-making process. As technology approached the present day, automated information systems became more capable and more involved in daily life. They began capturing more and more data, allowing the computer to become an active participant in expanding facets of daily decision-making. The exponentially increasing volume of available data has transformed the decision challenge from one of "data starvation" to "data saturation." Fayyad, Piatesky-Shapiro, Smyth, and Uthurusamy (Fayyad, et.al., 1996, pp. xv-xvi) attribute this "mountain of stored data" to such factors as advances in scientific data collection, introduction of bar codes, and the computerization of many business and government transactions. In many situations today, there is so much data that human beings are unable to correlate it all, and decision quality is again hampered, or in the words of John Naisbett (Fayyad, et.al., 1996, p. xv.), "We are drowning in information, but starving for knowledge."

Clearly there is a growing need for "intelligent agents," or automated information systems that can sift through these mountains of data (which other systems have efficiently collected) and integrate these sources into concise, usable knowledge for use in human decision-making. It is doubtful that a computer can reproduce the innovative creativity of a human analyst, but a computer system can be imparted with a basic representation of some of what the human analyst desires. This representation of interest is then used to filter vast volumes of available data (a task too time consuming for humans) and present the human analyst with a more concise body of knowledge in an understandable form. This premise is supported by many documents, such as this quote from Fayyad, et. al.:

Such volumes of data clearly overwhelm the traditional manual methods of data analysis such as spreadsheets and ad-hoc queries. Those methods can create informative reports from data, but cannot analyze the contents of those reports to focus on important knowledge. A significant need exists for a new generation of techniques and tools with the ability to intelligently and automatically assist

humans in analyzing the mountains of data for nuggets of useful knowledge. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD). (Fayyad, et.al., 1996, p. 2)

The Comprehensive Clinical Evaluation Program (CCEP) database presents this type of challenge to data analysis. The CCEP database contains vast amounts of information on over 19,000 Persian Gulf War (PGW) veterans who have brought some form of health concern to the attention of the Department of Defense (DoD) military healthcare system. The database contains a large number of attributes, and there are still no defined parameters for search. In any case, because of problem structure and sheer size, the entire database cannot be comprehensively analyzed by conventional means. The goal of this thesis is to design, construct, and implement an artificially intelligent computer system which can analyze the CCEP database more efficiently than a conventional or "brute force" approach without unduly taxing scarce medical research assets. Such computer systems are said to carry out "data mining."

B. PURPOSE OF THIS RESEARCH

The ultimate purpose of this research is provide the CCEP program with a viable methodology to obtain useful information from its database of participating PGW veterans. Determining what constitutes "useful" or "interesting" information is at least as great a challenge as devising an analysis tool. However, in the initial stages of medical research, interesting information is any statistical association between database attributes of different categorical groups. These associations may signal the existence of an undiscovered common ailment or "syndrome" affecting participants in the Persian Gulf War.

Time and other resources are also key factors in the overall CCEP research project. Simply investigating every possible combination of attributes may be theoretically feasible, but in actuality often necessitates an impractically large commitment of resources to the analysis task. Therefore, investigative speed and efficiency have become key factors in this research. The need for speed and efficiency demand that this research develop an intelligent search device capable of sifting through vast amounts of raw data and identifying interesting trends or correlations without the need for human intervention. Consequently, a genetic algorithm has

been selected. No commercial product suited our particular needs, so the purpose of this research includes the development and application of a genetic algorithm suited to analysis of medical data, specifically the CCEP database.

Finally, this research evaluated the success of the new genetic algorithm (DaMI, the NPS Data Miner) from several aspects:

- DaMI performance adheres to classical genetic algorithm theory
- DaMI statistical computations are valid and reproducible
- DaMI efficiently and comprehensively analyzes the search space
- Outcome hypotheses are of significant value to medical experts and the program sponsor

As with problem structuring, validation of results has proven to be a major research challenge and is addressed in this paper.

Computational results so far strongly suggest that our system has succeeded at the given tasks, requiring relatively few resources. They also have found no indication that a single syndrome or other medical entity is responsible for wide-spread adverse health ramifications among a significant cross-section of Persian Gulf War participants in the CCEP program. There are, however, numerous correlations of exposure/demographic information and associated symptoms/diagnoses which suggest that smaller groups may share common health conditions based on shared exposure to common health risk factors.

C. SCOPE OF RESEARCH

This research examines the problem structuring challenges for analyzing the data contained in the CCEP database. It discusses the general qualities of genetic algorithms and the specific techniques used to apply a genetic algorithm to the study of the CCEP database. The research focuses on application of a genetic algorithm to a relevant real-world problem and does not contain an in-depth description of genetic algorithm theory. An original genetic algorithm (DaMI) was created by this research effort. A technical description of the DaMI algorithm, its development process, and evaluation methodology are included. It is not the purpose of this

research to survey all possible solutions to the CCEP analysis challenge, but rather to completely examine and document one apparently successful solution. Finally, the results of the DaMI analysis of the CCEP database are presented along with the validation process and recommendations for further research. The following research questions were addressed:

- If there is a (actually there may be more than one) common ailment or “syndrome” afflicting veterans of the Persian Gulf War, how will it manifest itself within the scope of information gathered by the CCEP database?
- How will the subjective concept of interesting information (to the medical community) be quantitatively measured and used to compare the “fitness” of different hypotheses?
- How should the research problem and database be structured to facilitate automated analysis?
- Why is a genetic algorithm a more effective means of analyzing the CCEP search space than other more conventional methods?
- How was DaMI constructed? What were the design considerations and key innovations in this particular genetic algorithm?
- What analyses were conducted and what were the results?
- Were the results validated and were they useful to the project sponsor (CCEP, Deployment Surveillance Team) and CCEP medical researchers?

D. REAL WORLD APPLICABILITY

A great deal of research has been performed on genetic algorithms and related artificial intelligence-based research tools. In many cases, the data analyzed were real but in few cases the research was tied into a real world time-sensitive research problem. One of the primary reasons for using a genetic algorithm is that an answer is needed, but conventional research resources are not available to produce that answer within the allotted time. This makes a study of a real-world genetic algorithm development all the more interesting. The CCEP database research is highly-visible, relevant, and time-sensitive.

Only a select number of medical issues have received as much attention as the proverbial "Desert Storm Syndrome" in recent years. Since the first returning Persian Gulf War (PGW) veterans began reporting health issues, this subject has received constant attention by the U.S. government, military medical researchers, and most prolifically the media. A Presidential commission has been appointed to determine what, if any, health ailments may be attributed to the service of U.S. armed forces in the Persian Gulf. Research efforts continue at many DoD and Veterans Administration (VA) facilities. It is certainly appropriate to say that the CCEP is "high visibility."

Similarly, the concept of relating diseases to groups of humans with similar symptoms and life experiences (demographics and exposure to physical objects) has been a focus of medical research for many years. Some of the earliest genetic algorithm experiments attempted to relate symptoms to diagnoses. Medical science has consistently searched for better ways to answer the question, "What caused this disease?" In the case of CCEP, 697,000 veterans (not to mention their families) are eager to know if their service in the PGW increases their susceptibility to any type of medical malady. From an academic perspective, the issue of automatically identifying "interesting" information has become increasingly fascinating and challenging. Technology has increased researchers' ability to automate aspects of a medical situation, but the problem of making a model that accurately reflects the information remains.

E. THESIS METHODOLOGY AND ORGANIZATION

This research begins with examination of the CCEP research challenge as a whole. The first challenge is to structure the CCEP research question of what is an "interesting" hypothesis into a mathematical formula (fitness function). This in turn returns a higher "fitness" to hypotheses of greater interest to CCEP medical researchers. Our research tried many alternatives, but settled on the use of the Modified J-measure (described in section II.E.4.c) to assess relative independence between premise and outcome variables. The CCEP database was not designed with medical research in mind, so the second challenge was to reformat the database into a structure which supported automated analysis.

Once the problem and source database were structured appropriately, a suitable research tool was needed. It was clear that using a “brute force” approach to examine the CCEP database, even using computer simulation, was impractical because of the tremendous size of the search space. A genetic algorithm was chosen because of the innate ability of genetic algorithms to inductively adapt to the researcher’s goals and to intelligently analyze a search space, bypassing hypotheses which show little chance of future success. Our concept enhanced the conventional genetic algorithm approach by dividing the process into two modules: A genetic operator, which handles selection and recombination of hypotheses at the field level only, and a statistical package, which analyzes every possible combination of hypothesis fields passed from the genetic operator and returns an integrated fitness measure for the entire hypothesis. Additionally, our tool examines multiple independent and dependent (LHS and RHS) fields because CCEP could not determine which field or combination of fields would identify a target outcome.

Finally, the problem of validation and search space coverage must be addressed. A great deal of literature supports the idea that a genetic algorithm can deduce hypotheses that apply to a database. However, it is critical that these results be both validated against independent data and that they be indicated to accurately address the research question, instead of just exploring the data actual set analyzed. Several tools were developed to validate the results, among them an independent validation algorithm which independently re-tests results hypotheses against the subject database and a cross-validation procedure that tests hypotheses generated from one randomly-sampled subset of the databases against another randomly sampled subset.

The thesis is divided into seven chapters:

- Chapter I : Introduction
- Chapter II : Description of the CCEP Research, the database itself, and problem structure challenges
- Chapter III : Overall solution concept and high-level research approach
- Chapter IV : Description of the DaMI algorithm, its design, implementation, and validation processes
- Chapter V : Technical description of the DaMI algorithm operators, innovations, and procedures

- Chapter VI : Summary of results
- Chapter VII : Conclusion and recommendations for future research

F. ACKNOWLEDGMENTS

The author would like to acknowledge the financial support of the Assistant Secretary of Defense for Health Affairs, for allowing the purchase of equipment used in this thesis. This work was performed under Contract DRAM60040.

I would like to acknowledge several individuals whose contributions have, directly or indirectly, made this research possible. I am deeply grateful to all of them:

Professor Hemant K. Bhargava, provided a constant source of brilliant insight. He gave timely guidance but the freedom to learn. He has my highest respect as a academic, a teacher, and a human being.

LCDR Robert Glaser, MSC, USN, Head of the Comprehensive Clinical Evaluation Program, Deployment Surveillance Team provided the funding and support for this project when it was little more than a good idea. His constant support and "big picture" knowledge kept us in touch with the body of medical research efforts, which is so vital to any analysis effort.

LT Deb Lankhorst, USN, a decision systems colleague, acted as a sounding board, reality checker, and kept me from walking into trees as I looked at the big forest. Deb was the impetus for simultaneous analysis of the forward and reverse confidence and the production of rule text during genetic analysis.

The basis for my genetic algorithm theory comes from the work of John Holland. His landmark research into the use of genetic theory for knowledge space search provides the basis of DaMI operating principles. My basic genetic algorithm operators are direct applications of the work of David Goldberg. His clear presentation of basic genetic algorithm operators in (Goldberg, 1989) were the starting points for DaMI.

DaMI represents its hypotheses as strings and is therefore a genetic algorithm and not genetic programming. In spite of this, I found Professor John Koza's (1990)

descriptions of genetic programming tunable parameters, common pitfalls, and representation of real-world problems to be extremely helpful. I used his work as the basis for selection of tunable parameters and developed my random-fill same-parent crossover technique as a genetic algorithm solution to the premature convergence problem. He solved the same problem using genetic programming crossover.

II. COMPREHENSIVE CLINICAL EVALUATION PROGRAM

A. BACKGROUND AND HISTORY OF CCEP

The Department of Defense (DoD) began to examine the health consequences of Persian Gulf War (PGW) service while U.S. troops were still deployed to the Persian Gulf Region. The initial focus of medical researchers was on the health risks associated with smoke from Kuwaiti oil fires. As early as 1992, groups of PGW veterans began presenting with health complaints which they attributed to PGW service. Many of these veterans reported nonspecific symptoms or those not directly attributable to a specific disease or syndrome (group of commonly occurring symptoms/conditions). This sparked the first of many tests (first by the Army in 1992 and subsequently by other services) to attempt to discover if these non-specific symptoms could be linked with any "clusters" of PGW veterans. The theory of this approach is that a new syndrome will present as a "cluster" or group of individuals sharing some common trait (demographics, location, action, exposures, etc.) who also share a similar group of symptoms. (CCEP, 1996, pp. 6 - 7) This is the first step to identifying a new syndrome. Once a syndrome is defined, then medical researchers begin efforts to find the cause of the syndrome. If a solid cause-effect relationship is established and documented between an entity (virus, bacteria, etc.) or health risk factor(s) (like smoking or cholesterol), then the syndrome may be considered a full-fledged disease.

In response to the health concerns of PGW Veterans, both DoD and Veterans Affairs (VA) established similar comprehensive clinical evaluation programs. The data for this research comes from the DoD CCEP. The CCEP program was officially enfranchised by the Assistant Secretary of Defense (Health Affairs) as part of a three-point plan, announced on 11 May 1994. This plan included:

- *The development of an aggressive, comprehensive, clinical diagnostic program to offer intensive examinations to veterans who do not have clearly defined diagnoses,*

- *An initial independent review of DoD clinical and research efforts concerning the Persian Gulf War by Dr. Harrison C. Spencer, Dean of the Tulane School of Public Health and Tropical Medicine, New Orleans, Louisiana, and*
- *The creation of a forum of national medical and public health experts to review, comment, and advise DoD concerning the results of the clinical evaluation program.* (Joseph, 1994)

CCEP continues to offer in-depth medical examinations, through the Military Health Services System (MHSS) to any PGW veteran having health concerns. Over 27,000 PGW veterans and their dependents have initiated medical examinations with CCEP, of which over 19,000 have been completed by the participants. The data collected from these 19,000 participants has been recorded in a single database (the CCEP database), which is the source database for this research. (CCEP, 1996, pp. 7 - 12)

Since the inception of CCEP, numerous medical research programs have been conducted by DoD and non-DoD health organizations (including the Defense Science Board, National Institute of Health, Naval Health Research Center in San Diego, University of California, Department of Health and Human Services, and National Academy of Sciences). Although several research efforts are still ongoing, the possibility of an unknown syndrome or disease affecting PGW veterans and their families has been exhaustively examined. DoD has committed to continue research on this issue but stated:

To date, there is no clinical evidence for a previously unknown, serious illness or 'syndrome' among Persian Gulf veterans participating in the CCEP. A unique illness or syndrome among Persian Gulf veterans evaluated through the CCEP, capable of causing serious impairment in a high proportion of veterans at risk, would probably be detectable in the population of 18,598 patients. However, an unknown illness or a syndrome that was mild or affected only a small proportion of veterans at risk might not be detectable in a case series, no matter how large. (CCEP, 1996, p. 4)

It is this viewpoint that has catalyzed the need for an intelligent, automated search program to analyze the CCEP database. Clearly, conventional research (user-controlled query and clinical evaluation) has reached the limit of available resources, and yet there is still a possibility that a syndrome has remained undetected. Proper implementation of a genetic algorithm can expand

the horizon of research by sifting through hypotheses not yet considered but will do so using small amounts of time, funds, and human effort.

B. CCEP RESEARCH VISION

The core of CCEP research is based on classic epidemiological technique. The CCEP database has been constructed to capture as wide a range of data about PGW participants as is practical. Data collection practices have been standardized and unbiased--any participant with a concern undergoes the same health screening and examination process. The basic premise of analysis is that a new syndrome will present as "prominent and consistent physical and laboratory findings" like Legionnaire's disease or toxic shock syndrome or consistent "non-specific symptomatology" as with chronic fatigue syndrome and fibromyalgia.

In any case, CCEP research efforts focus on slicing the database in many different directions, whether by demographic information, symptoms, diagnoses, or reported exposure categories. Percentages of PGW participants in each slice or "cluster" (which is a group of participants with the same characteristics within a given research slice) are compared to the percentage expected within a similar population not participating in the PGW. In many cases (especially when the database is sliced by reported exposures), no comparable group is available, so these percentages are compared against actual percentages or distributions among all 697,000 PGW personnel (as opposed to just those participating in CCEP). The point of the analysis is to isolate any characteristic which appears to make a CCEP participant more likely to have approached CCEP with a medical condition.

If some specific combination of demographics, personal habits (smoking/non-smoking), and reported exposure is associated with specific symptoms and diagnoses with the group of CCEP participants, then medical research is developed to clinically test the relationship of these factors to personal health. It should be apparent that this approach is extremely resource intensive. Analysis dimensions are limited to the imagination of individual researchers developing the slices and the physical ability of medical researchers to examine the hypothesis. If the quality of "statistical interest" could be mathematically modeled by an automated research tool, then the dimensions of analysis could be expanded to the limits of computer (as opposed to

human) resources. The genetic algorithm (DaMI) is a research tool designed specifically to relieve humans from the drudgery of human-controlled analysis so that they may focus efforts on clinical testing which machines cannot do.

C. DATABASE DESCRIPTION

The CCEP database is a "flat file" or single table with 177 attributes. It was created in standard dBase® format and was actually received and manipulated using the Visual Foxpro® Database Management System (DBMS). The database was not designed with automated analysis or medical research (for that matter) in mind. Therefore, a great deal of manual file manipulation was required before automated analysis was possible. By "manual" we mean the issuance of single SQL® commands to reformat individual database schema and field values. At no time was the actual data adjusted, but in many cases the representation schema was changed to enhance automated processing. Appendix A contains the CCEP data dictionary alone, a commentary on modifications/usability of each field, and a synopsis of the CCEP data collection process. The actual database used for research contains 17,033 records for active duty CCEP participants. Dependent records were removed prior to analysis at the request of the CCEP program manager.

A large number of attributes containing administrative and/or privacy act data were removed from the database and other attributes were added to enhance the schema, as discussed above. (For a more complete description of schema modifications, see section II.D.2) In all, 140 attributes were present in the research database. Not all were examined at once (see Section VI.A), but in any case the database was relatively large by medical or occupational health research standards. The remaining attributes fall into four major categories:

- **Demographic.** Physical attributes of each participant (e.g. race, gender, age, home state, service component, Unit Identification Code [UIC])
- **Reported Exposures.** Reported exposures to potentially hazardous environmental conditions by participants (e.g. botulism vaccine, oil smoke, uranium, passive smoke, local water, SCUD attack)

- **Reported Standard Symptoms.** Standard symptoms elicited by physicians during CCEP medical examinations (e.g. difficulty breathing, fatigue, headaches)
- **Diagnoses.** Each participant completing the entire CCEP medical examination process was assigned a primary and up to six secondary diagnoses. Diagnoses followed the standard numeric ICD coding system (e.g. V65.5 - Healthy Exam, 307.81 - Chronic Muscle Tension Headaches, 780.71 - Fatigue)

As will be seen in later sections, most analysis was conducted on associations between these major attribute categories.

D. WHY DOES A GENETIC ALGORITHM WORK FOR CCEP ANALYSIS?

1. Theory

The theory of genetic algorithms was invented by John Holland in the early 1970's. Holland's purpose was to create a search method based on the process of natural selection observed in nature. He likened the attributes making up a hypothesis in a search problem to chromosomes which "encode" a living being. He proposed that by creating mathematical representations of genetic reproduction and applying natural selection, scored by a fitness function, to those representations, he could create an adaptive search engine. Automation of this process has proven to be an excellent task for computer systems. Although a great deal of evolution is not understood, several general features are agreed upon: (Davis, 1991, pp 2 - 3)

- Evolution is a process that operates on chromosomes rather than on the living beings they encode.
- Natural selection is the link between chromosomes and the performance of their decoded structures. Processes of natural selection cause those chromosomes that encode successful structures to reproduce more often than those that do not.

- The process of reproduction is the point at which evolution takes place. Mutations may cause the chromosomes of biological parents, and recombination processes may create quite different chromosomes in the children by combining material from the chromosomes of two parents.
- Biological evolution has no memory. Whatever it knows about producing individuals that will function well in their environment is contained in the gene pool--the set of chromosomes carried by the current individuals--and in the structure of the chromosome decoders.

If one is to follow the theory of natural selection, then it could be inferred that attributes used to make hypotheses are the operators of evolution. The process of hypothesis evolution revolves around the combination of those constituent attributes of successful hypotheses and their resulting recombinations. Furthermore, these recombinations are directed blindly and guided only by the principle that attributes belonging to hypotheses of higher fitness measure are recombined more frequently than attributes belonging to hypotheses possessing lower fitness measure.

Holland went on to create three genetic operators which could mathematically recombine the modeling chromosomes of coded hypotheses to mimic genetic recombination. Hypotheses from the gene pool of the current are "selected" with a bias towards hypotheses with higher fitness measures, and then operated on by one of these three genetic operators:

- **Reproduction.** Asexual reproduction of single parent rule to single offspring rule without modification
- **Crossover.** Sexual reproduction involving the exchange of chromosomes between two parents producing two different child rules.
- **Mutation.** Asexual reproduction of single parent rule with random modifications resulting in a different child rule.

Using the "Two-armed and k-armed bandit problems," (see Holland, 1975 for complete proof) Holland went on to prove that, lacking prior knowledge of the expected value of two or multiple

choices, allocating slightly more than exponentially increasing trials to choices with the highest past success is the optimal means for choosing between options. The results of this theory and its relation to genetic operators is summed up well by Goldberg:

In other words, to allocate trials optimally (in a sense of minimal expected loss), we should give slightly more than exponentially increasing trials to the observed best arm...Another method that comes even closer to the ideal trial allocation is the three-operator genetic algorithm discussed earlier. The schema theorem guarantees giving at least an exponentially increasing number of trials to the observed best building blocks. In this way the genetic algorithm is realizable yet near optimal procedure (Holland, 1973a, 1975) for searching among alternative solutions. (Goldberg, 1989):

It is important to reiterate that genetic algorithms gain their speed, not by analyzing an entire search space, but from deciding which attributes (chromosomes) hold the least probability of producing interesting hypothesis and not testing hypotheses using those attributes. The process is not fixed, for it relies on probability for modeling, and different results will be derived each time the algorithm is run. This fact will be discussed further in the discussion of results validation.

Now let's bring this theory closer to the current research question. A hypothesis concerning the CCEP database may be "encoded" into a string representing its constituent attributes. If one is to hold with Holland's theory, then the attributes (in this case demographic, exposure, symptom, or diagnosis) which make up the hypothesis (in a group or hypotheses) having the highest fitness measure should be recombined in an exponentially increasing number of fashions. Similarly, the attributes from unsuccessful hypotheses should be recombined exponentially less often. Genetic operators, used in the DaMI genetic algorithm, prove be the most optimal way of accomplishing this selection. Finally, if this process is followed, then the extremely large search space of correlations within the CCEP database will be searched most efficiently using a genetic algorithm. It is on this theoretical basis that we chose a genetic algorithm to analyze the CCEP database.

2. Advantages and Disadvantages of the Genetic Algorithm Method

There is a great deal of theoretical literature on the advantages and disadvantages of using genetic algorithms. It is the intent of this section to relate practical lessons learned from our specific research using DaMI on the CCEP database. From the point of view of this research, a genetic algorithm was particularly useful because of its ability to process tremendous amounts of data and its lack of need for human interaction. It has already been proven that CCEP problem search space is too large to analyze by conventional means, even with a computer. The problem cannot be structured strongly enough to limit the possibilities to realistic numbers, so technology is being relied upon to perform the discrimination. Medical research assets are a scare resource, so employing medical experts only at the fitness function creation and final analysis stages produces efficient and effective results. Should preliminary implementation of genetic algorithms prove informative in this area of medical research, many other similar research questions may benefit from this technology.

There are several disadvantages to using genetic algorithms, several to which have already been alluded. First, as can be seen from section II.D, a great deal of effort must be committed to database structure and normalization before processing. Since the system relies on computer evaluation of data, the data structure and coding scheme must be uniform and conducive to information extraction. Non-descriptive representations and textual data collection will severely curtail system performance. The strong coding and standardization of the CCEP database was one of the aspects that made it so attractive for this type of research. Second, a genetic algorithm is useless without a single, unambiguous representation of what is interesting to the operator. This was a key challenge to this research. There are many measures which may infer the "interestingness" of a particular hypotheses, but the synthesis of a single aggregate measure which satisfies all components of epidemiological interest has been extremely difficult (several different fitness functions may be required). Finally, a difficult paradox arises when attempting to *prove* that a genetic algorithm has completely searched a large space. A genetic algorithm achieves its speed advantage by selective analysis, meaning it selectively eliminates search options with, apparently, little chance of yielding interesting results. The only way to

actually prove that an interesting hypothesis was not missed is to physically test every hypothesis, but we turned to the genetic algorithm because the resources necessary to search the entire space were not available. To address this problem, the genetic algorithm is run several times. If the outcomes produced by several independent runs have a high intersection (particularly among hypotheses of high fitness), then there is strong evidence that the space has been searched adequately. A more detailed discussion of this challenge is included in Chapter V.

To sum up, this research has found that genetic algorithms do search a very large space of alternatives very quickly and efficiently. Successive generations of hypotheses quickly improve in quality as measured by the fitness function, and therefore the algorithm does adjust its search to the operator's goals. Strong database standardization and coding are a must before any processing is attempted. A genetic algorithm has proven successful to this research, as long as a fitness function can be created which accurately defines "what is interesting" to the researchers.

E. KEY CHALLENGES TO CCEP ANALYSIS BY A GENETIC ALGORITHM

1. Problem Structure

The single most challenging aspect of this research is that "Persian Gulf Syndrome" as it is referred to by the media, PGW veterans, and some researchers, is not yet really a defined syndrome at all. A syndrome must be defined by a unique series of symptoms and/or ailments which are shared by a specific group of individuals. Although many PGW veterans report a wide array of non-specific medical ailments associated with PGW service, no defined set of symptomatology has been enstantiated as a candidate syndrome.

CCEP clinicians have identified a wide range of specific diagnoses (i.e. migraine headache, depression, asthma, arthritis, hypertension). However, few if any of the conditions diagnosed to date could be considered specific for any of the many different exposures implicated as potential causes of Persian Gulf illnesses. Thus as a case series, the CCEP has identified a wide spectrum of different clinical conditions rather than any singular homogeneous diagnostic entity (CCEP, 1996, p. 79)

While the medical implications of this statement are serious, the impact of this situation on research is tremendous. Basically, CCEP medical researchers cannot provide us with a description of a target syndrome for research, or for that matter if there are one, many, or any syndrome(s) at all. Without target syndrome characteristics, a researcher is unable to identify which field or combinations of fields within the database indicate a desired outcome (a syndrome of interest). In truth, researchers do not know if the data necessary to identify a syndrome, should one exist, is contained in the database at all. Therefore, we have been compelled to develop a tool which can examine “interesting” associations between any number of causative and outcome attributes without specificity as to the limits of either the causative or outcome space. This is both a curse and a blessing; the lack of specifics makes the problem considerably more challenging but also stimulates interest in our type of tool.

What can be reasonably asked about the problem is the following:

- **Is there a syndrome?** Is there subset *a* (of *A*) ailments such that the occurrence rate of *a* in PGW participants (*G*) is higher than the rate in a reference population (*R*)? [*#a*(*G*) equates to “number of occurrences of an ailment within the set of participants (*G*)”]

$$\frac{\#a(G)}{\#(G)} > \frac{\#a(R)}{\#(R)}$$

- **What caused the syndrome?** Is there a subset *x* (of *X*) of exposures and/or demographic experienced/attributed to participants in the PGW such that: for ailments *a* for which the prior equation is true, exposures/demographics *x* account for a significant part of the difference in occurrence rates of *a* in groups *G* and *R*?

$$P(a|x, G) = \frac{\#a(G)}{\#x(G)} \neq \frac{\#a(R)}{\#x(R)} = P(a|x, R)$$

The lack of precise target syndrome definition encourages the development of multiple research strategies. As mentioned before, the directed query technique used by CCEP (CCEP, 1996, pp. 17 - 49) has sliced the database from numerous different perspectives. What is needed is a search tool which can examine multiple combinations of independent (LHS) and dependent (RHS) variables and all possible values for each variable simultaneously. This adds an extra dimension to the analysis. Conventional data mining tools typically allow the user to specify a range of possible LHS variables for search and a single RHS variable. Multiple RHS fields may still be handled under this doctrine by creating a pseudo field which contains a different value for each unique combination of values in the RHS fields to be examined. However, if the RHS fields for analysis are large in number or cannot be specifically identified, the pseudo field coding becomes impractically large. What is needed instead is a data mining tool which can apply selective induction operators to a range of possible attributes (not just individual attribute and value instances) on the LHS and RHS simultaneously.

This methodology is plausible and in fact was done by DaMI in this research, but it is prudent to note that this strategy will still produce an extremely large search space. For example, the first analysis done by DaMI examines the associations between 15 standard symptoms (LHS) and 21 possible diagnoses (RHS). All attributes are Boolean and are not limited in the number of simultaneous combinations (all symptoms and diagnoses could be simultaneously present or "true"). Therefore the possible search space is 2^{36} or 6.8×10^{10} possible hypotheses. It is for this specific reason that we chose to use a genetic algorithm, with its ability to discriminately analyze tremendous search spaces. A test was conducted in which this particular problem was analyzed using simple "brute force" (test every possible combination indiscriminately), using a 486DX/66 Mhz personal computer. The personal computer was able to test about 600,000 combinations per day. At this rate, this one complete analysis would take 114,992 days (315 years). Even if a platform were chosen that was 100 times faster than our test personal computer, the analysis duration would be an unacceptable 3.15 years.

2. Database Content and Structure

Several problems were encountered during the course of this research with the CCEP database content and structure. These problems fall into two major categories: data representation anomalies which make it difficult for an algorithm to extract meaningful information from the data, and data collection anomalies which introduce bias into the data being analyzed. Examples of data representation anomalies include irrelevant data and non-normalized data. These problems must be corrected before useful analysis can be conducted; they usually require modification of the database itself. In the case of CCEP, data collection anomalies include data that were self-reported by participants, self-referral of PGW veterans to the CCEP program, and lack of an established control group. Collection anomalies do not interfere with analysis itself, but they must be acknowledged or accounted for when examining results.

Seventy-seven fields in the CCEP database are simply unusable. Many fields contain sensitive unclassified data on the participants (names, social security numbers, addresses, etc.) which is not helpful for medical research and is subject to the Privacy Act of 1974. Those fields were deleted at the outset. Another larger group of fields is used by CCEP for administrative processing and are similarly not helpful to research. Finally, there were some fields that have been collected as non-standardized text. The most serious occurrence of this is the "chief complaint" or in other words the reason that the participant approached CCEP for an examination. No standardization was enforced in this free-text field so it is relatively impossible for a computer to determine similarity between tuples, short of creating a complete index of chief complaint texts and some standard category indicator. This is fortunately not the case with diagnoses, which use the standard numeric ICD coding system. Participant complaint information was captured in the form of fifteen standard symptoms, but a coded chief complaint would prove most helpful.

A key shortcoming of the database, reported at the outset by CCEP, is the large amount of data which are self-reported by participants. Self-reported data are that which is directly determined by responses from participants during their medical examinations (as opposed to clinical test results, review of documentation, or impartial third-party observation). Self-reported data are analogous to a survey, which is in and of itself not a database flaw. However, in the

context of CCEP, all exposure and standard symptom data are self-reported. This reduces the direct applicability of aggregate participant responses because perceived exposure may be distinctly different from actual exposure. This is most easily demonstrated by an example we call “the Botulism Illusion.” Within the CCEP database, 26.4% (4,500) of the active-duty participants report receiving the botulism vaccine. Now it is known from medical records that only 8,800 or 1.26% of the 697,000 PGW veterans were given this vaccine. This high percentage (26.4% of participants) would appear to suggest a possible relationship between the botulism vaccine and PGW medical ailments, until it is pointed out that 21.9% of the CCEP participants who were examined and deemed “healthy” (primary diagnosis of V65.5) also reported receiving the botulism vaccine. (See Figure #1) Problems concerning *reported* data may be compensated for by collecting and examining a “control group” of participants who do not have significant medical conditions; however, reported data should always be interpreted with some degree of caution.

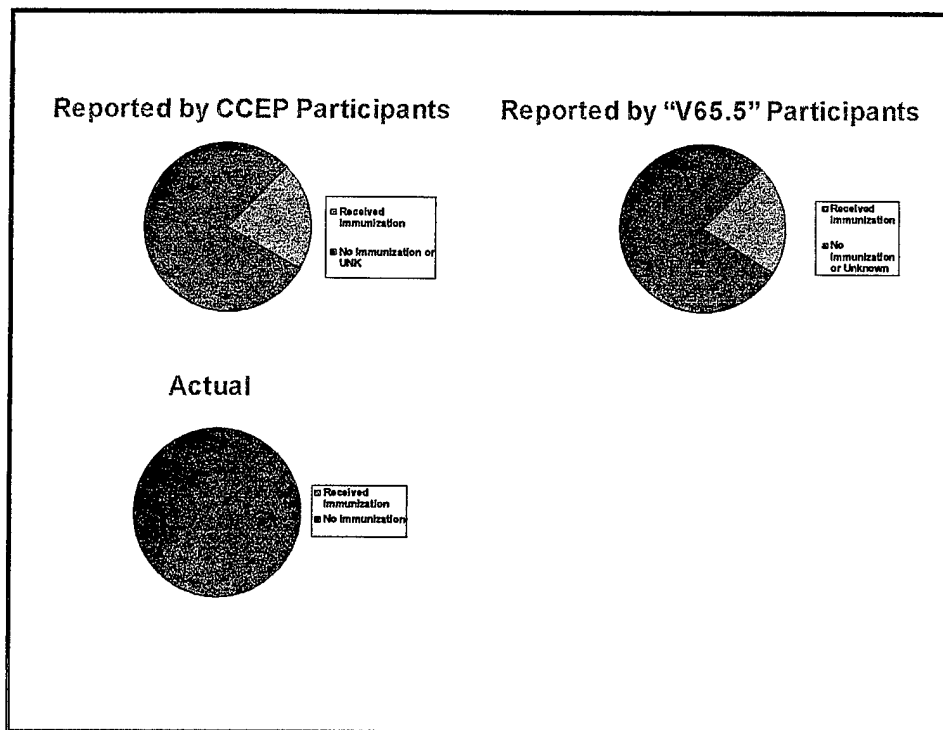


Figure 1. The Botulism Illusion

Another obstacle to a meaningful analysis of the CCEP database is the self-referral (participants made a conscious decision to start the CCEP examination process) of participants. As described in Appendix A, any individual who was eligible for medical care under the MHSS system in 1994 and had a health concern related to PGW service (whether directly or indirectly) could request a full medical evaluation under the CCEP program. This encouraged a wide range of participants, but the self-referral of patients may invalidate the CCEP database as a statistical representation of PGW veterans as a whole. Had the participants in CCEP been selected randomly, then their aggregate response and demographic data could have been considered statistically representative. In this case, the sheer act of self-referral introduces some level of bias which, if it can be identified, should be explained to the degree possible. One possible solution is to randomly select a suitably large group of PGW veterans, regardless of health concerns, and provide them with the same medical evaluation as the other, self-referred, participants. In other words, create a control group. A control group will help identify bias from both self-reporting and self-referring. Unfortunately, this has not been adopted as part of the CCEP program. Suggestions have been made to create a control group after-the-fact, but a strong argument can be made that the passage of time since 1994 will introduce similar bias into the responses of a present-day control group.

The reader should not infer that the CCEP database is a poor source; it has many strong points. After removal of unusable fields and reformatting other fields for enhanced analysis, 140 "good" fields have remained for analysis. One of the most positive aspects of the database, is the standardization of CCEP data collection. From the outset, CCEP used the same database structure, examination process, and coding scheme for all medical examinations. There are some exceptions, such as the case of chief complaint (mentioned above) but overall the data content is strongly coded and standardized. Any reader who has dealt with data analysis at all, should appreciate the importance of a uniform database structure and coding system to computer analysis. Something as simple as representing an affirmative response as "Y" or "Yes" or "yes" can make computer-based query far more difficult. Of particular significance was the uniform usage of numeric ICD codes to represent outcome diagnoses.

3. Database Normalization

The uniform coding scheme used in the CCEP database and limited need for scalar (continuous numerical) data sharply reduced the need for normalization (when used in a data mining context, "normalization" means structuring a database for effective computer analysis). The coding scheme used in the CCEP database is quite strong, so only a few modifications were made to normalize the database. Three significant modifications were made to the schema for analysis. Diagnoses were converted from single fields to multiple Boolean fields to facilitate analysis of diagnosis combinations. Standard symptoms were changed from durations to simple occurrence to simplify the ambiguity of comparing duration categories. Finally, an aggregate reproductive disorder field was created to relate reported reproductive disorders of any type.

a. Boolean representation of diagnoses

The CCEP database captures outcome diagnoses assigned by the examining physician as a primary diagnosis and six secondary diagnoses. CCEP researchers assign a somewhat higher emphasis to the primary diagnosis, and place little weight on the ordering of secondary diagnoses. Therefore, a medical researcher would not differentiate between a diagnosis of fatigue appearing second or say fourth on a list of diagnoses attributed to a participant. A computer on the other hand could consider these distinctly different occurrences. Since combinations are tantamount to this research, it is much easier to represent and analyze a string of diagnosis fields with Boolean (yes or no) operators than a string of up to seven unordered diagnoses. However, 1700 different diagnoses were assigned to the 19,000+ CCEP participants, so a pure Boolean representation would be extremely unwieldy. We decided to represent the twenty-one most frequently occurring diagnoses as Boolean operators in addition to the existing ICD representation. The number twenty-one was selected arbitrarily (it can be expanded in future research), but at least one of the selected diagnoses is included in 74.7% of participant outcomes. See Figure #2 below.

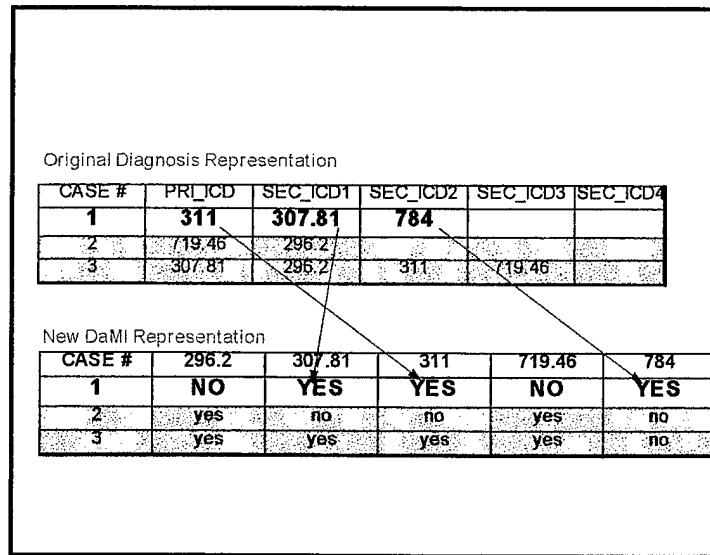


Figure 2. Diagnosis Attribute Restructuring

b. Standard Symptoms

In the CCEP database, participants are asked to report suffering from fifteen standard symptoms (e.g. chest pain, difficulty breathing, head aches). The responses are collected dates of onset and duration. The date and duration are subjective (and subject to error), and like diagnoses, difficult for an automated search engine to compare. A higher confidence can be assigned to a response if it is represented as a Boolean (the participant will in most cases accurately report existence of the symptoms, while his/her ability to estimate an onset and duration is questionable). Therefore, fifteen additional fields are added to the CCEP database, one corresponding to each symptom and equal to "Y" if the participant reported the symptom at any time for any non-zero duration.

c. Reproductive Disorders

One of the high visibility aspects of the PGW is the possibility that a syndrome may be causing PGW participants to experience a higher rate of reproductive disorders (specifically birth defects). The CCEP database captures reproductive disorders (participant may

report reproductive disorder actually experienced by a spouse or manifested in offspring) in five areas:

- Infertility
- Miscarriages
- Still births
- Infant deaths
- Birth defects

These five categories are further subdivided into disorders experienced prior to and after PGW service, making a total of 10 reproductive disorder fields. We cannot be certain that a syndrome, should it exist, would cause only one form of reproductive disorder. Therefore, two new fields were created to reflect any reproductive disorder experienced by the participant, either prior to or after the PGW conflict. In other words, if a participant reported infertility, a miscarriage, a still birth, an infant death, or a child with birth defects prior to PGW service, then the new field (PQ_prior) was set to "Y." If none of these were experienced prior to PGW service, then PQ_prior was set to "N." Similarly, if any of the five sub-categories were affirmatively answered after PGW service, then PQ_after was set to "Y." This will allow the research to be more sensitive to associations between demographic, exposure, symptom, and diagnosis data and any combination of reproductive disorders. Naturally, any interesting associations developed concerning these two new fields will need to be re-categorized by medical researchers before a finding may be made.

After completion of normalization, 6 demographic, 32 reported exposure, 15 (Boolean) standard symptom, and 21 (Boolean) diagnosis fields are available for automated analysis. These 74 fields observe a uniform structure and coding scheme and are the foci of this research. Please consult Appendix A for a detailed list of analyzed fields.

4. What is “Interesting?”

In Section II.D.1, we asked the question, “What is a syndrome?” It is necessary at this point to revisit this question, but from an automated analysis perspective. A genetic algorithm depends (as do many other techniques) on the ability of the researcher to define in quantitative terms what is “interesting?”. The problem in many forms of decision science is not whether a model performs accurately, but rather if it improves the quality of a decision. In a genetic algorithm, selection of hypotheses to evaluate is proportionally related to a “fitness” value for each hypothesis, so it is critical that our “fitness function” accurately represents the interest of medical researchers. This characteristic is reflected in the fundamental genetic theory:

“Roughly, the fitness of a phenotype is the number of its offspring which survive to reproduce...This measure rests upon a universal, and familiar, feature of biological systems: Every individual (phenotype) exists as a member of a population of similar individuals, a population constantly in flux because of the reproduction and death of the individuals comprising it. The fitness of an individual is clearly related to its influence upon the future development of the population. When many offspring of a given individual survive to reproduce, then many members of the resulting population, the “next generation,” will carry the alleles of that individual.” (Holland, 1975, p. 12)

This returns us to the fundamental question: “What is interesting to CCEP medical researchers and how will that interest be manifested in the database?” In Section II.D.1, we stated that we are not sure whether a syndrome exists, and, if it does exist, we are not certain that the data captured in the CCEP database are appropriate to identify it. However, if these two uncertainties are removed, the following assertions can be made:

- If there are one or more syndrome(s) affecting PGW veterans, the data to identify them may already exist in the CCEP database but is hidden by the sheer volume of data.
- In this case, a syndrome will manifest itself as a single or unique group of diagnoses or symptoms shared by a cluster of participants sharing some common exposure and/or demographic attribute(s)

By plunging directly into a search for associative relationships between risk factors and outcomes, we bypass a fundamental step in classical epidemiological technique. Normally, epidemiologists will first define the outcome diagnoses and/or symptomatology which describe a prospective syndrome. Once the definition is made, then research efforts are focused on associations with risk factors and other exposure sources. Unfortunately, the present research is left with a less than optimal situation. We suggest that a promising use for a genetic algorithm is to give clues to medical researchers that help them define a syndrome.

In this research, we have accepted that conventional research methods alone may not be able to define and isolate a syndrome affecting PGW veterans. We are now led to re-examine the problem from different perspectives. Our research approach has been guided by the following ideas:

- We are not trying to create an analysis that will isolate a single pre-defined Desert Storm Syndrome. Instead we are defining a profile that a syndrome might follow, should it exist. Our goal is to determine how a possible syndrome would be reflected in the data, as discriminately as possible, and then construct a fitness function which is appropriately high when this profile is met.
- Our genetic algorithm does not find a Desert Storm Syndrome, but rather distills the billions of possible hypotheses into a set of hundreds. All in the set of candidate hypotheses are not syndromes, but if a syndrome(s) does(do) exist, it(they) will be found in the candidate set. This smaller set of candidate hypotheses may realistically be examined more exhaustively by medical researchers and other conventional means.
- By implementing the genetic algorithm as a precursor to medical research (and alleviating the idea that it must find "the answer"), we allow the genetic algorithm to significantly reduce the burden on the relatively scarce medical research assets at a relatively small cost to the organization. In more basic terms, the secret to operating genetic algorithms in an imperfect world is to allow them to do the first 80% of the analysis work with only 20% of the research cost.

With the question of “interest” now bounded, a proper fitness function may now be pursued. If a true syndrome does exist, then it is “caused” by something. Therefore, the participants will share some finite set of exposure mediums, or in other words all participants with a syndrome will share some commonality in exposure. This must be caveated by saying that the CCEP database may or may not contain the demographic and exposure elements to identify that commonality of exposure. But as our research mindset states, we are only attempting to establish the profile of a syndrome if it exists, and if the data necessary to identify it is contained in the CCEP database. If the prior statement is true, then there will be a relatively strong association between a finite set of exposure/demographic attributes and a unique combination of outcome diagnoses. Likewise, there will be a strong association between a finite set of exposure/demographic attributes and a specific combination of standard symptoms. The intersection between diagnoses and symptom combinations with similar exposure associations will profile a candidate syndrome. See Figure #3 below.

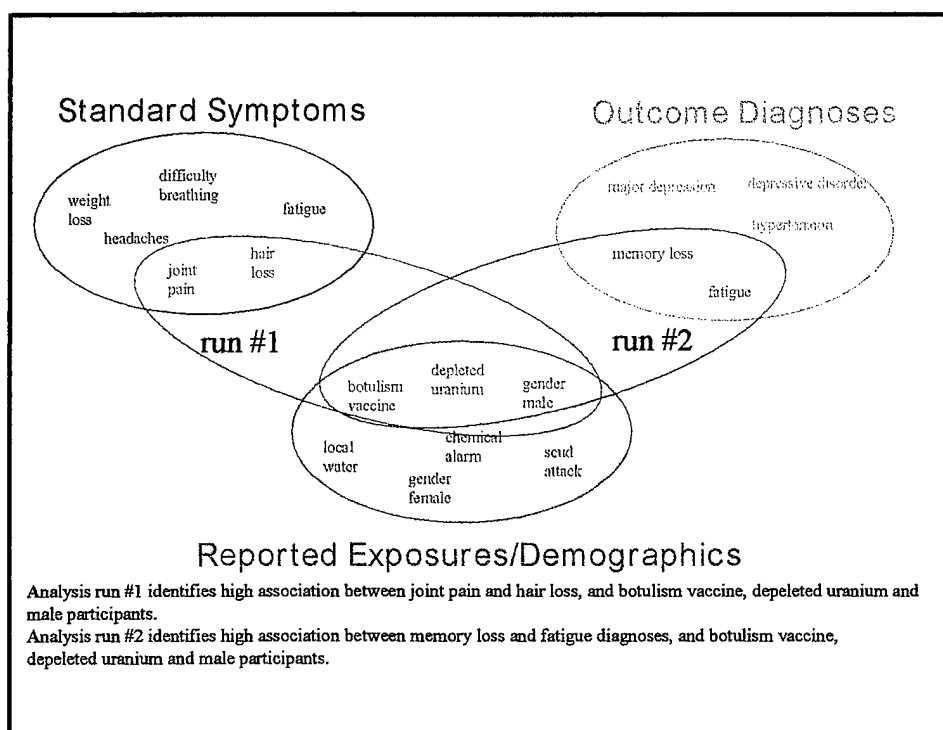


Figure 3. Hypothesized Syndrome Profile

Now our question of “what is interesting?” can be defined. “Interesting” is combinations of RHS attributes (dependent variables) which are highly dependent on combinations of LHS attributes (independent variables), or in other words, the candidate dependent variables are truly determined (not independent of) by the candidate independent variables. The fitness function used must be such that hypotheses which demonstrate this property will be assigned a relatively high fitness value. There are numerous accepted functions in statistical literature that fit this requirement. Several of these are discussed in the next section.

a. Conventional Epidemiological Measures

A great deal of literature already exists, like (Goldberg, 1989) and (Holland, 1975), to support the idea that genetic algorithms are quite successful at adaptively improving the quality of tested rules to suit the provided fitness function. From the outset, our genetic algorithm demonstrated this quality. However, the greatest challenge has been to ensure that the search model adequately represents the research questions (i.e. the genetic algorithm is doing what it was told to do, but have we provided it with relevant, meaningful instructions?). As a starting point for development of the fitness measure for this research, we first turned to classical epidemiology literature.

Classical epidemiology evaluates any test in terms of four variables (see Figure #4 below) which describe how successfully a test predicts the actual presence (or lack) of a specified disease. This is much akin to our own research which attempts to identify the success of a single or multiple exposure and/or risk factor attributes predicting a combination of symptoms or clinical diagnoses. In epidemiology, these four variables {a, b, c, d} are computed using a two-by-two matrix of test results and actual disease presence.

		Disease		
		Present	Absent	
Test	Positive	a True Positive	b False Positive	$PV(+)$ $a/(a+b)$
	Negative	c False Negative	d True Negative	$PV(-)$ $d/(c+d)$
		<i>Sensitivity</i> $a/(a+c)$	<i>Specificity</i> $d/(b+d)$	

Figure 4. Classical Epidemiological Measures

By mathematically manipulating these four variables, four “quality” values are obtained from the relationship between the subject test and subject disease. In each case, keep in mind that our research is applying the risk/exposure as a test for (or indicator of) a specific symptom and/or diagnosis profile. These quality values are (Fletcher, 1982, pp. 43 - 57):

- **Positive Predictive Value.** Indicates the ability of a positive test result to accurately identify the presence of a disease in a patient. This term is similar to “confidence” used as a fitness measure in many data mining tools. We term this “forward confidence.”

$$PV(+) = \frac{a}{a+b}$$

- **Negative Predictive Value.** Indicates the ability of a negative test result to accurately determine the absence of a disease in a patient. Most data mining tools do not consider this measure, but recommend the analysis be run with swapped dependent and independent variables. This is not practical if multiple dependent variables are being analyzed.

$$PV(-) = \frac{d}{c+d}$$

- **Sensitivity.** The proportion of subjects with a disease who have a positive test for the disease. A sensitive test will rarely miss people with the disease.

$$sensitivity = \frac{a}{a + c}$$

- **Specificity.** The proportion of subjects without the disease who have a negative test. A specific test will rarely misclassify people without the disease as diseased.

$$specificity = \frac{d}{b + d}$$

b. Fitness Measure Paradoxes

In our research, classical epidemiology measures are helpful in choosing a suitable fitness function, but no single aforementioned measure is sufficient for several reasons. Rather we desire an aggregate fitness measure which will increase in response to any classic measure of interest. Fundamentally, this research problem differs from clinical test evaluation in one respect. While a high number of either false positive (b) or false negative (c) tests is a counter-indication of a test's quality, it is also desirable (in our case) if a risk/exposure combination is contraindicative of an outcome symptom/diagnosis set. In certain cases, a true positive may mean nothing because there are also many false positives. In other cases, a simultaneously high false positive and false negative is quite informative. This is best described by an example (Figure #5), but basically, in the case of CCEP database analysis, we are most interested in the hypotheses having highest values and lowest values of sensitivity and specificity.

→ Consider the most simple hypothesis, 1 LHS (L) and 1 RHS (R) field.

- If L and R are Boolean, there are four possible hypotheses to test.
- We are looking for more than just a high $\text{prob}(R=\text{"yes"}|L=\text{"yes"})$.

INTERESTING		NOT INTERESTING	
IF L = "yes" THEN R = "yes"	90%	IF L = "yes" THEN R = "yes"	10%
IF L = "yes" THEN R = "no"	10%	IF L = "yes" THEN R = "no"	90%
IF L = "no" THEN R = "no"	80%	IF L = "no" THEN R = "no"	80%
IF L = "no" THEN R = "yes"	20%	IF L = "no" THEN R = "yes"	20%

∞ As the number of fields and/or values per field increases, the problem expands exponentially

Figure 5. Attribute Value Relationships

c. *Alternative Fitness Measures*

Now that our concept of "interesting" has been framed from the epidemiological perspective, we can set about the task of selecting a single fitness measure which mathematically describes our concept of interest to the genetic algorithm. Again, there is some challenge in this because there are several different measures of interest to medical researchers (discussed in the previous section), yet the genetic algorithm requires a single aggregate fitness measure. The genetic algorithm could be run several times using different fitness measures, but this carries a high cost in both processing time and post-processing analysis effort. Likewise, we have seen from the preceding section that reliance on any single measure carries with it the possibility of statistical misinterpretation. Two paths were examined in this research to address this problem, although we note that there may be many other possible solutions.

- **Modified J-measure.** Refer again to Figure #4 and the four test characteristics [PV(+), PV(-), sensitivity, and specificity]. Our first approach was to create a measure which was suitably large when any of these four measures were large and suitably low when none of the measures were relatively large—in effect an aggregate fitness measure. It should be noticed from the foundation we have laid that if both a

and *d* are relatively large when compared with *b* and *c*, the four test characteristics are all relatively large. This would demonstrate that the risk factors and/or exposures under investigation are highly successful in predicting the outcome symptoms and/or diagnoses under investigation. Tentatively we will select the following formula as our fitness measure:

$$mod_j(fitness) = \frac{a \times d}{b \times c}$$

It may also be noticed that this measure will effectively indicate if the outcome symptoms/diagnoses are successful at predicting the risk/exposures. We call this property, "reverse confidence." It is particularly helpful to examine the two sets of attributes with each assuming the role of dependent and independent variables simultaneously. Finally, recall that unlike the evaluation of clinical tests, CCEP analysts consider it interesting if both false positive and false negative values are simultaneously high (indicating a risk/exposure combination reduces the probability of a symptom/diagnosis combination). To account for this situation, our *j*-measure is modified as follows

$$\begin{aligned} \text{if } \left(\frac{a \times d}{b \times c}\right) \geq 1, mod_j &= \frac{a \times d}{b \times c} \\ \text{if } \left(\frac{a \times d}{b \times c}\right) < 1, mod_j &= \frac{b \times c}{a \times d} \end{aligned}$$

(Figure #6 gives an example of a modified *j*-measure calculation; note we use a natural log function to shape the fitness function for better genetic competition; this will be discussed in Chapter V):

$$\text{mod j-measure} = 1 + \ln[(a*b)/(c*d)]$$

$$1 + \ln(11*7505)/(84*146) = \mathbf{2.91}$$

Fatigue

	“yes”	“no”	
Uranium Exposure	a 11	b 84	<i>PV(+)</i> $11/(11+84)$ $= 11.6\%$
	c 146	d 7505	<i>PV(-)</i> $7505/(146+7505)$ $= 98.1\%$

Sensitivity
 $11/(11+146)=7.0\%$

Specificity
 $7505/(84+7505)=98.9\%$

Figure 6. Modified J-measure Calculations

- **Chi-square.** Another approach to the question of fitness function may be derived strictly from statistics. Since our aim is to identify risk factors and/or exposures that are highly associated with symptom and/or diagnoses groups, we may use a statistical principle which measures the independence (not the same as the term "independent variable" used in knowledge discovery science to denote the RHS variables) of two groups of attributes. According to Walpole, et. al, "The chi-square test procedure...can also be used to test the hypothesis of the independence of two variables of classification." (Walpole, et. al., 1988, pp. 343 - 346) The same "contingency table" used by epidemiologist, may be constructed and used to compute expected levels of **a**, **b**, **c**, and **d** based on the joint probability function of the dependent and independent variables. (See Figure #7) Observed values are the original values of **a**, **b**, **c**, and **d**, and expected values are calculated using the following formula:

$$\text{Estimated_Expected_Value} = \frac{(\text{column_total}) \times (\text{row_total})}{\text{grand_total}}$$

The chi-square is now calculated and summed for all cells in the matrix. (*Chi-square may be used for any size matrix, in this case two were used for simplicity. Since a two-by-two matrix is used in the example, the formula below contains the Yates Correction, which is not necessary in larger matrices.*) A higher chi-square indicates a higher level of dependence (or *lack of independence*) between the two attribute sets. The Chi-square formula (with Yates correction) follows; example chi-square calculations are included in Figure #7 :

$$\chi^2 = \sum_i \frac{(|o_i - e_i| - .5)^2}{e_i}$$

		Fatigue		
		"yes"	"no"	
Depleted Uranium Exposure	"yes"	a 11(1.93)	b 84(93.07)	95
	"no"	c 146(155.07)	d 7505(7495.93)	<u>7651</u>
		157	7589	7746

Figure 7. Chi-square Calculations

The modified j-measure has been used by this research to date, however a new statistical analysis package designed to analyze using chi-square is currently being constructed. A more straightforward formula for Chi-square will actually be used in the new statistical analysis package (Dixon and Massey, 1969, pp. 242 - 243):

$$\chi^2 = \frac{(|ad - bc| - \frac{1}{2}N)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

III. SOLUTION CONCEPTS

A. RESEARCH GOALS

In the case of the Desert Storm research, years of conventional medical research have yielded no single syndrome or associated symptomatology set. This means that the no fixed dependent variable set (combinations of diagnoses and/or reported standard symptoms) can be readily identified. The traditional epidemiological paradigm is to isolate a group of individuals with consistent symptoms/outcome diagnoses and then find what key demographic or exposure elements these individuals share. If relating demographic/exposure data are present, it is used to focus clinical research on an underlying cause. This approach has not proven fruitful to date, either because no syndrome exists or because the sheer volume of data in the CCEP database hides a relation of interest from human-controlled querying. Therefore, we have chosen to let technology simplify the problem from the outset of the knowledge discovery process.

As mentioned before, there are four basic categories of useful data contained in the CCEP database {demographics, reported exposures, reported standard symptoms, and outcome diagnoses}. While attributes in each category could prove useful as independent (LHS) or dependent (RHS) variables, it is doubtful that attributes from the same category will be useful as both LHS and RHS simultaneously. The research question is now simplified to an examination of which attributes (or combinations of attributes) in each category are most highly associated with (or statistically dependent on) which attributes from another major data category.

EXAMPLE *What associative relationships exist between exposure attributes and outcome diagnosis attributes? Based on analysis, there is a high association between reported exposure to Scud Attack and Depleted Uranium and an outcome diagnosis of Post-traumatic Stress Disorder. [This is just an example, not an actual finding]*

This exponentially increases the size of prospective search space which is represented by $2^{\#LHS} * 2^{\#RHS}$ (where #LHS = number of independent fields and #RHS = number of dependent

fields and all attributes are Boolean; if not the search space is even greater). The increase in search space can provide useful insight to medical researchers as they develop hypotheses. Instead of waiting for medical researchers to provide a more structured problem (and thereby reduce the search space), it was our feeling that an intelligent search technique could be employed effectively in the problem as given. Therefore, the role of our genetic algorithm is to test an extremely large subset of all fields in the CCEP database concurrently for levels of interest based on a specific model of epidemiological interest, to wit:

$$\theta(LHS^*, RHS^*) = \max(\theta(LHS', RHS'))$$

where $LHS' \subset LHS^*$ and $RHS' \subset RHS^*$ and $\theta() = \text{fitness function}$

We did count on CCEP medical researchers to define their concept of "interesting" and thereby guide our selection of an appropriate fitness function. This fundamental shift in knowledge discovery technique suggests that a genetic algorithm may be used to provide researchers with information to assist them in framing the initial research strategy, instead of framing the problem and then passing it to a genetic algorithm. We asked the following question, "If a syndrome does exist and the data necessary to identify it are contained in the CCEP database, what data relationships would it create in the CCEP database?" The answer to this was converted to a mathematical fitness measure. The resulting combinations of exposures/demographics and symptoms/diagnoses discovered will contain any identifiable syndromes', but the entire set of hypotheses will not all be guaranteed to be useful solutions. The goal is to present medical researchers with a more workable solution space in which to focus their conventional research efforts. This approach shifts the burden of searching a tremendous alternative space appropriately onto the genetic algorithm.

B. SOLUTION STRATEGY

Our solution strategy takes two forms, theoretical and practical. In the theoretical sense, the solution strategy rests on selection of the most efficient method of searching an extremely large solution space. There are three basic methods of search:

- **Random.** In this type of search, a computer program will randomly generate hypotheses and pass these hypotheses to an evaluating routine. The evaluating routine assigns a fitness measure to each hypothesis based on the fitness function provided. If the hypotheses are generated sequentially, this method is also known as "brute force." This method tests many hypotheses, because the hypothesis generation apparatus is extremely simple, but has no capacity to self-improve or tune the search to the operator's goals.
- **Human-controlled Selective Search.** In this case, a human formulates a hypothesis and translates it into the form of a query. The query is evaluated by the computer system and the results are returned to the human operator. It is assumed that the human operator draws upon practical knowledge of the problem and the results or prior queries to formulate new queries. Therefore, the quality of query formulation improves throughout the process. This allows the search to self-improve (*including the human operator within the boundary of the search system*) and obviously tune to the operator's goals. However, the hypothesis generation is extremely slow.
- **Systematic, Intelligent, Automated Search.** A computer program (genetic algorithm) generates hypotheses, passes them to an automated evaluator, receives results, and then re-generates a new set of hypotheses (*systematically adapting its search based on its past performance as indicated in the results received*). This technique demonstrates all three desirable search characteristics: fast hypothesis generation, self-improvement, and tuning to the operator's goals.

Figure #8 illustrates the comparative advantages of each search technique. It should now be clear, from a theoretical point of view, why a (genetic algorithm) systematic, intelligent, automated search has been chosen.

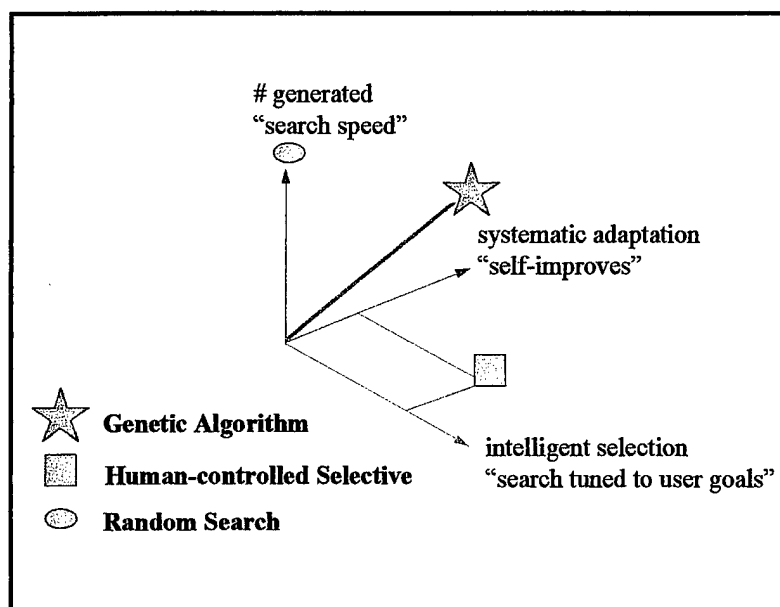


Figure 8. Characteristic of Different Search Techniques

Now let us discuss the solution strategy on a more practical level. Assume for a moment that a genetic algorithm performs a systematic, intelligent search as theorized. The next section will provide a theoretical basis for this assumption. From Section II.D.4, we draw the premise that a syndrome will manifest itself as a high association between a specific combination of demographic and/or exposure attributes and a finite set of symptomatology or diagnoses. Combine this with premise that either a modified j-measure or chi-square formula will indicate the level of association (or dependence) between two sets of attributes. Our strategy is then to instruct the genetic algorithm (DaMI) to find the most significant associations between demographics/exposures and symptoms and between demographics/exposures and diagnoses. These two analyses will divide the complete set of possible combinations of demographics/exposures into three categories (note that demographics/exposures are traditionally viewed as the independent attribute set):

- **Demographic/Exposure combinations which appear on neither analysis.** Any hypothesis not contained on either study indicates that there is no statistical basis within the CCEP database to indicate that combination is a possible syndrome. This does not mean that it could not suggest a syndrome; as stated before, the CCEP database may not capture the appropriate data to identify the hypothesis as a syndrome.
- **Demographic/Exposure combinations are associated with both specific combinations of symptoms and specific combinations of diagnoses.** This is the ideal case for suggesting the existence of a syndrome. It indicates that a group of PGW participants, sharing both a common symptomatology and outcome diagnosis set belong to the demographic profile and/or report common exposure elements. Clinical research should be directed toward a prospective syndrome demonstrating the listed symptoms and diagnoses. Again this indicates that a hypothesis meets the mathematical definition of interesting, but the possibility of it being a syndrome can only be confirmed by evaluation by medical professionals.
- **Demographic/Exposure combinations are associated with either specific combinations of symptoms or diagnoses.** A majority of hypotheses identified by DaMI will fall into this category. If only one correlation is made with the demographic/exposure data, there is a weaker indication that this particular combination signals a candidate syndrome. However, failure to appear on both analyses should not completely discount the hypothesis. As mentioned before, the failure of the CCEP database to capture all symptomatology or diagnoses may explain the appearance of the demographic/exposure combination on only one analysis. Therefore, hypotheses in this category should still be evaluated by medical professionals.

Naturally, a certain degree of ambiguity exists concerning the specific fitness measurement thresholds with respect to interest (filtering). Filtering will be discussed in Chapter VI. But in a practical sense, this analysis will provide medical researchers with a prioritized list of interesting associations. The central point is that most possible hypotheses will prove statistically

implausible and therefore fall into the first category, suggesting they not receive costly conventional medical research efforts.

Finally, many initial DaMI discovery sessions were devoted to analyzing relationships between reported symptoms and outcome diagnoses. Early input from CCEP epidemiologists included a strong desire to identify unexpected symptom/diagnosis combinations. This study was appealing for initial research because all attributes involved were Boolean (as opposed to demographic and exposure attributes having more than two possible values). The research proved statistically successful (discussed in Chapter VI) but of limited practical value to CCEP.

IV. DaMI GENETIC ALGORITHM ARCHITECTURE

Up to this point, this thesis has focused on the theoretical structuring of the CCEP research problem and formulating the qualities of a genetic algorithm required to solve the problem. The second half of this thesis will focus on describing the tool developed to meet these challenges and the success of that tool in actual analysis. Based on the preceding discussion, the genetic algorithm must be specifically designed:

- to accept an unstructured set of dependent and independent variables
- efficiently search an extremely large search space
- employ adaptive learning, where a priori information is used to guide future hypothesis testing

This chapter will deal with DaMI from a macro systems perspective; Chapter V will address the details of the system's design.

A. PROGRAM MODULES

Unlike many other genetic algorithms, the system designed for this research (DaMI) has been using several independent modules. These modules consist of the genetic algorithm itself, a statistical package, a user interface, and a verification package. There were two primary reasons for this design strategy. The first was to relieve the genetic algorithm of the mundane analysis tasks, results filtering, and user interface tasks, thereby enhancing the space searching efficiency. The second reason was to aid in system development. By adopting a modular development approach, a great deal of effort can be focused on the core genetic algorithm technology and allow the system to begin rapid prototyping before optimal statistical analysis and user interface modules were developed. Once the core genetic algorithm is properly functioning, more robust statistical engines and user options may be added, using experience gained from test runs. A

more in-depth explanation of the genetic algorithm (GA) operation is contained in the next chapter. Figure #9 shows the relationship between the DaMI modules.

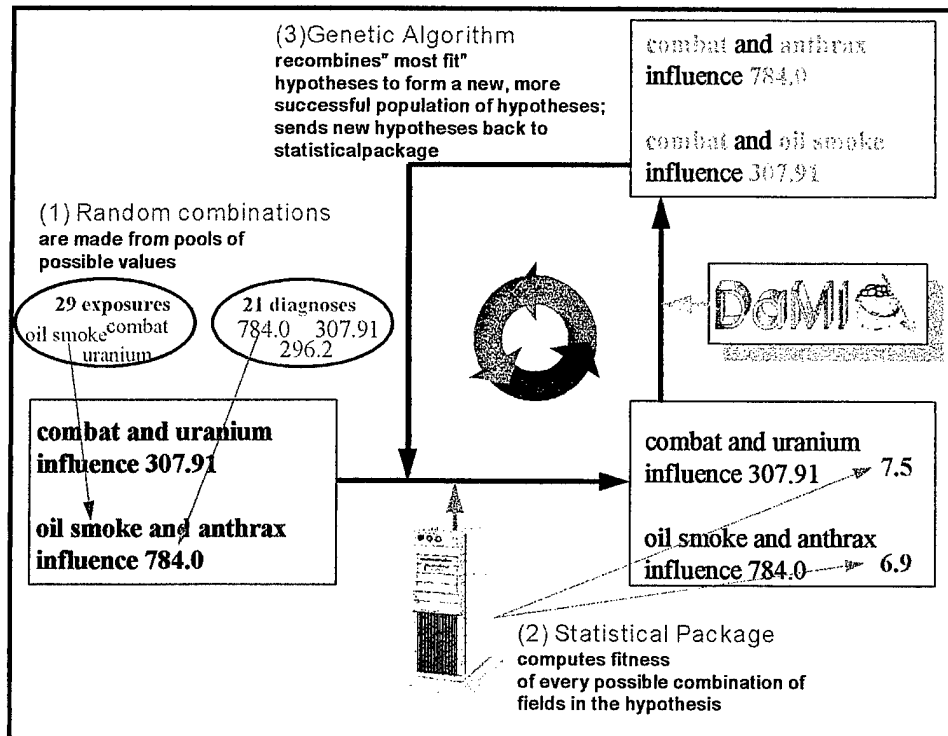


Figure 9. Relationship of DaMI Modules

1. The Genetic Algorithm Package

The genetic algorithm package is responsible for maintaining a list (population) of hypotheses (rules) in the current generation, selecting the most successful rules, and performing the genetic operations of reproduction, crossover, and mutation. These genetic operators allow the system to adapt the analysis to the goal model (fitness function) and improve the search hypotheses as each generation is processed. In this thesis, "hypothesis" and "rule" are used interchangeably; "hypothesis" is a medical research term and "rule" is a artificial intelligence term. Clearly, not all possible hypotheses will be tested (hence the advantage of the genetic algorithm), but the use of genetic operators ensures that the rules being tested have the highest probability of satisfying the given fitness function (Holland, 1975). In the DaMI system, the genetic algorithm stores hypotheses as combinations of attributes only, not as combinations of

attributes and specific values. Competition is based on success of attribute sets as a whole. Attribute sets (like gender, receiving the botulism vaccine, exposure to uranium [independent variables] and Depression and Chronic Fatigue Syndrome [dependent variables]) are passed to the statistical package, which returns an aggregate fitness value for all possible value combinations of those attributes. The statistical package is called recursively during the processing of a single generation for every rule, until the entire generation is evaluated. Then the genetic algorithm produces the next generation and the process is repeated.

2. The Statistical Analysis Package

The statistical analysis package receives a set of independent and dependent attributes to evaluate from the genetic algorithm package. The statistical package requires no information other than a list of field names to evaluate. The number of attributes in each request sent to the statistical package varies, so it must be capable of processing loosely bounded problems. During pre-processing, the analysis database (database under analysis; in this case the CCEP Persian Gulf War Database) is examined and a table is created of all attributes and their possible values. This table is used as the source for generating each individual query (there are many individual queries generated to answer each request from the genetic algorithm) and ensuring that each possible combination is tested but only once. The statistical package then computes the fitness of each possible attribute/value combination. An aggregate fitness measure is then computed and returned to the genetic algorithm package. As the statistical package tests attributes against the database under analysis, it also performs a test of each attribute/value combination against a second database. This second test is not returned to the genetic algorithm and therefore does not affect hypothesis competition. This value is stored to be used later for results validation (see section V.C).

3. User Interface

The user interface controls interaction between DaMI and the system operator. The user interface allows the user to adjust tunable parameters (discussed in Chapter V), view the discovery database at various stages of processing, and start and reset the genetic algorithm package. The user interface also provides intermediate feedback to the user during DaMI operation. It was designed using the Foxpro Screen Design Wizard and is controlled by push buttons and pop-up menus. Settings may not be adjusted “on-the-fly” when the genetic algorithm is operating. An example of the user-interface screen is shown in Figure #10 below. The user-interface module is disposable, and therefore an in-depth discussion of the user-interface design is not included in this thesis.

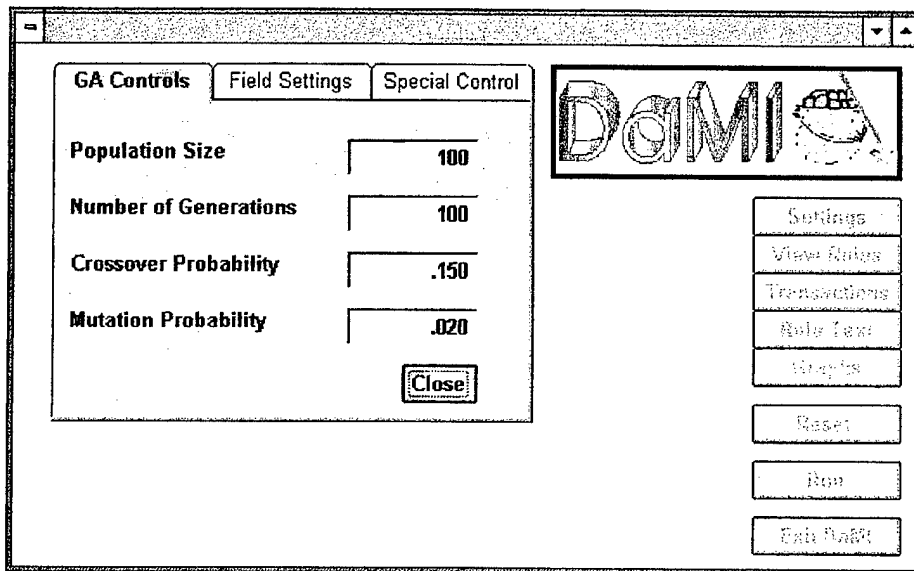


Figure 10. DaMI User Interface

B. REPORTING AND FILTERING

Once a discovery session has been completed by DaMI, several files are created. A transcript of each hypothesis individual (at the attribute level) of every generation is created as DaMI operates, along with a transaction record of each genetic operation employed, the source (parent) rules, and resulting offspring. The transaction record also maintains a time stamp at the start of each generation which can be used to monitor processing speed. DaMI also records how many actual combinations were tried during the session. These files will not be discussed in detail (file structures are contained in Appendix B).

The most important file created (rulelib.dbf) contains a list of every hypothesis tested and used to determine an aggregate fitness measure (without duplication). Several key points must be cleared up at this juncture. First, not every possible attribute/value combination is used to compute the aggregate fitness value of a given attribute set (this is a tunable parameter). Second, Rulelib.dbf stores attribute and value combinations (as opposed to the session transcript which records only the higher-level attribute sets). It also contains the intermediate, final, and verification fitness measures. This makes rulelib.dbf the actual answer produced by DaMI. Figure #11 is an excerpt from rulelib.dbf.

Rule number	No true th	No true rh	No true bo	No false b	Standard c	Reverse cf	Complex cf
12	23	625	3	7101	0.13	0.00	1.54
3	2052	840	205	5059	0.10	0.24	1.12
22	320	268	6	7164	0.02	0.02	1.65
7	122	464	4	7164	0.03	0.01	1.64

Vcomplex	Lhs_text	Rhs_text
0.49	SERVICE="X"	A780_9="Y"
1.03	CHEM_ALARM="N"	A780_71="Y"
1.38	SERVICE="4"	A780_7="Y"
1.11	HEAT_SMOKE="Y".and.MUSTRD_GAS="Y"	A477_9="Y"

Figure 11. Rulelib.dbf Display

Finally, whatever fitness measure is used will probably not have an arbitrary threshold of “interest.” A fitness measure is only useful in ranking the relative interest of hypotheses tested; therefore some form of filtering will be done prior to reporting. However, it is inadvisable to enforce that filter during operation. Instead, rulelib.dbf is left in the most robust (non-summarized) form practical; filtering is performed arbitrarily using SQL type query language on a case-by-case basis for each report.

Several reports have been developed in Foxpro for the DaMI system. However, as with filtering, reports are tailored to suit the needs of each individual recipient. Summary reports are created on an ad-hoc basis; there is a standard detailed report which contains hypotheses and all intermediate and final statistical computations. The detailed reports (two main studies were conducted in this thesis) of the top 100 hypotheses discovered are contained in Appendix C.

C. SYSTEM REQUIREMENTS

1. Hardware and Software Requirements

From the outset, the author's goal was to construct a research tool and methodology that can be employed by researchers in their community, without the need for a laboratory of (scarce) high-power computer assets. In any case, it has already been shown that raw processing power is quickly overcome by large unstructured database analysis requirements. Therefore, a genetic algorithm is used to intelligently enhance the processing capabilities of whatever platform it runs on. In keeping with this goal, DaMI was designed to operate on a standard personal computer using inexpensive commercial software. The hardware and software requirements required to run DaMI are listed below:

Hardware Requirements

Personal Computer, 80486/66Mhz processor or better

8 Megabytes of RAM

200 Megabytes of free hard disk storage

Software Requirements

Microsoft® Visual Foxpro version 3.0

Microsoft® Windows version 3.xx or Windows 95

Surpassing the minimum hardware requirements will of course benefit system performance. The most dramatic performance improvements will be realized by increasing RAM and the access speed of the PC hard drive.

2. Processing Limits

DaMI is primarily limited by the time available to the user to complete the analysis; however, there are some processing limitations. For the preservation of system speed, DaMI maintains the active population in a RAM-based array. Therefore, it is limited by the maximum array size allowed in Foxpro. The required array size is a function of population size per generation and number of attributes under analysis. The formula for this metric is:

$$population_size \times analysis_fields \leq 73,500$$

Under this limitation, analysis of 70 field with a population size of 15,000 (array size 1,050,000) would exceed the system limits. Only the number of fields actually under analysis is used in this calculation, not the number of fields in the database being analyzed. Also, the number of records in the analysis database is limited only by the maximum Foxpro table size (Maximum records per table file = 1 billion, Maximum size of a table file = 2 gigabytes, Maximum fields per record = 255). Naturally, larger files will take longer for the statistical package to analyze.

V. SEARCHING THE HYPOTHESIS SPACE: DaMI IMPLEMENTATION

A. THE GENETIC ALGORITHM

The basic architecture of the DaMI Genetic Algorithm is based on (Goldberg, 1986), with the notable exception that our genetic algorithm stores rules as strings of Boolean attributes ("true"=*consider the attribute*; "false"=*don't consider the attribute*). This allows the genetic algorithm to process simple binary strings, as opposed to strings of field values and wildcards (Goldberg uses a "*" to denote any value of this attribute is acceptable). This does not imply that the genetic algorithm is simplistic, in fact competition of attributes in aggregate actually provides for a more efficient search of the alternative space. As can be seen in Figure #12, a conventional genetic algorithm will operate hypotheses as combinations of attributes and values. In our case, this prevents the genetic algorithm from considering the associations between risk factors (exposures/demographics) and outcomes (symptoms/diagnoses) in aggregate. By using the DaMI methodology, risk factors and outcome associations (hypotheses) are examined comprehensively before competing for selection and genetic recombination.

Conventional Genetic Algorithm Representation (Goldberg, 1989)										
	Demographics		Reported Exposures				Outcome Diagnoses			
Rule	Gender	Service	Uranium	Oil	Smoke	Combat	Anthrax	Fatigue	Depression	Memory Loss
1	male	Navy	Yes	*	*		No	*	Yes	*
Rule 1 indicates a relationship between Male Navy personnel who reported exposure to Uranium but not Anthrax and an outcome diagnosis including Depression										
DaMI Genetic Algorithm Representation										
	Demographics		Reported Exposures				Outcome Diagnoses			
Rule	Gender	Service	Uranium	Oil	Smoke	Combat	Anthrax	Fatigue	Depression	Memory Loss
2	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
Rule 2 indicates a relationship between gender, service, reported exposure to Uranium and/or Anthrax and whether or not the patient was diagnosed with Depression										

Figure 12. Conventional and DaMI Algorithm Representations

This genetic algorithm uses a "roulette wheel" (Goldberg, 1989) model for competitive selection with the size of each rule's "slice" (or probability of selection) being directly proportional to the fitness measure (determined by the statistical package) of each rule. Slices are selected for reproduction, crossover, and mutation randomly, but the "size" of each slice gives a proportionally higher chance of survival to rules with higher fitness. As individual rules show reproductive dominance, these individuals may possess more than one slice on the roulette wheel. (i.e. a particularly strong rule may reproduce more than once per generation, giving it more than one slice on the subsequent generation's roulette wheel). We chose the roulette wheel (Goldberg, 1989) because it allows the stronger rules to dominate more quickly than with other methods (e.g. rank or tournament) and thereby converge faster. The basic genetic operators (reproduction, crossover, and mutation) are all implemented in DaMI, with operator adjustable profiles (see section V.D).

B. THE STATISTICAL ANALYSIS ALGORITHM

The DaMI statistical package in use is a fairly simple algorithm. The modular design of our system allows for the replacement of this statistical package with a more robust commercial package in the future. At this point, the cost of designing an interface outweighs potential benefits; this may not be true for more complex analysis projects.

Given a set of dependent attributes (RHS) and independent attributes (LHS), the statistical package creates a two-dimensional array of attributes and possible values. The array also contains the number of possible values for each attribute and a counter for each attribute. As the statistical algorithm processes each combination, the counter for each attribute is incremented accordingly using the base counting of each attribute corresponding to that attribute's number of possible values. (i.e. if the attribute "gender" had two possible combinations then its counter would increment in base 2; if the attribute "state" had fifty combinations then its counter would increment in base 50). The algorithm uses each individual attribute's current counter value to reference a cell in the array. The cell values and attribute names are used to create a textual query statement. The query statement is then applied to the analysis database and the fitness measure is applied to the result. This allows the same statistical algorithm to loop recursively with a minimum amount of software code, regardless of the number of attributes passed to it by the genetic algorithm.

Several fitness measures have been used (see the discussion in section II.E.4). Our goal, since medical researchers seek associations between patient risk factors/exposures, reported symptoms, and resulting diagnoses, is to award the highest fitness values to those LHSs and RHSs which are most highly interdependent (vice independent). Since each request from the genetic algorithm generates many individual statistical package queries, some means of aggregating the fitness measures of all possible combinations is required. Several different methods for determining the aggregate fitness measure were considered. Obviously, an average of all fitness measures for a given attribute set is non-competitive. In many cases, the highest individual fitness measure has been used because of the specificity of the research question. In other cases, an aggregate measure may be taken using Chi-square or an average of the top three

or four j-measures (use of an aggregate value limits the awarding of a high fitness measure based on a single unexpected outlier in the research database).

A rule cacher (like a disk cacher, except for hypotheses) is used to prevent duplicate evaluation of any rule throughout the discovery session. A table of rules evaluated by the statistical package and resulting fitness values is maintained. Before sending a rule to the statistical package, the genetic algorithm checks the table of rules already evaluated. If the rule has been previously evaluated, the genetic algorithm uses the fitness value from the cache table. If not, the genetic algorithm package sends the rule to the statistical package and establishes a new entry (with resulting fitness) in the cache table.

C. TUNABLE PARAMETERS

The program has several tunable parameters to adjust genetic algorithm operation. Tunable parameters are set via the user interface at the commencement of each discovery session.

- **Crossover probability.** probability that a selected rule will exchange information with another selected rule
- **Mutation probability.** probability that a selected rule will undergo a random mutation
$$prob(reproduction) = 100\% - (prob(crossover) + prob(mutation))$$
- **Population size.** number of individual rules in each generation number of generations to simulate
- **Maximum rule complexity.** maximum number of dependent and independent attributes allowed in each hybrid rule (set individually for dependent and independent)
- **Average complexity of initial rule set.** average number of dependent and independent attributes allowed in each rule of randomly generated initial population
- **Top rules to aggregate.** number of rules (in order of decreasing fitness) to use in computing aggregate fitness by the statistical package

D. PROBLEMS AND IMPROVEMENTS

Before this discussion of DaMI implementation is concluded, we would like to discuss some of the problems encountered in our implementation and our solutions to these problems. We found, as many other researchers have, that genetic algorithms are quite successful at adaptively improving the quality of tested rules to suit the provided fitness function. However, the greatest challenge has been to ensure that our search model adequately represented the research questions (i.e. the genetic algorithm is doing what it was told to do, but have we provided it with accurate instructions). Our focus on problems with proper tuning of the genetic algorithm should in no way degrade the perception that a genetic algorithm is an extremely fast and effective search technique. It does work as advertised!.

1. Convergence Issues

One challenge faced by our research was to ensure that the algorithm would effectively (not necessarily physically) test the entire search space. A genetic algorithm will rapidly (especially using roulette wheel competition) improve the average fitness measure of rules within successive generations, but in many cases, the speed of improvement degraded the algorithm's ability to comprehensively examine the search space.

It should be recalled from genetic search theory (Holland, 1975) that search regret (or missed rules of interest) is minimized if attributes of successful rules are tested in exponentially more combinations in successive generations, and attributes of unsuccessful rules are tested exponentially fewer times. This is implemented in a genetic algorithm by giving successful rules a higher chance of selection (and thereby the chance to mix information with other successful rules) based on the level of their fitness measure. Naturally, successful rules begin to dominate the population (in our case take up more slots on the roulette wheel) and increase the chance that their constituent attributes are used for future rules. A problem arises when the fitness measure of a mediocre rule is disproportionately larger than the other individuals of its generation. If this mediocre rule dominates the population too quickly then its attributes provide the only material

for future rules. The resulting phenomenon is called *premature convergence* (Koza, 1988) and will prevent comprehensive search of the entire space.

Several steps were taken to prevent this, but generally speaking, great care must be used in selecting a fitness measure. If the slope of fitness in proportion to rule quality is too great, premature convergence is likely. The author chose to apply a natural logarithm scale to the fitness measure. This gave a strong relative advantage to good rules over weak rules, but slowed the domination of good rules (or local maximums) over their slightly weaker peers. The author also developed a technique called *same-parent crossover randomization*. Basically speaking, if two identical parents are selected for crossover, the resulting "offspring" are duplicates of the parents. In our crossover operator, if the two parents are the same, a single parent is randomly bisected into two offspring. Each offspring receives a portion of the parents genetic material (attributes) and a portion of randomly generated material. This has no effect on the algorithm at early stages, but it increases the mutation probability strongly as the population becomes dominated by a few rules (which causes the crossover operator to lose its ability to effectively generate new hypotheses, see Figure #13).

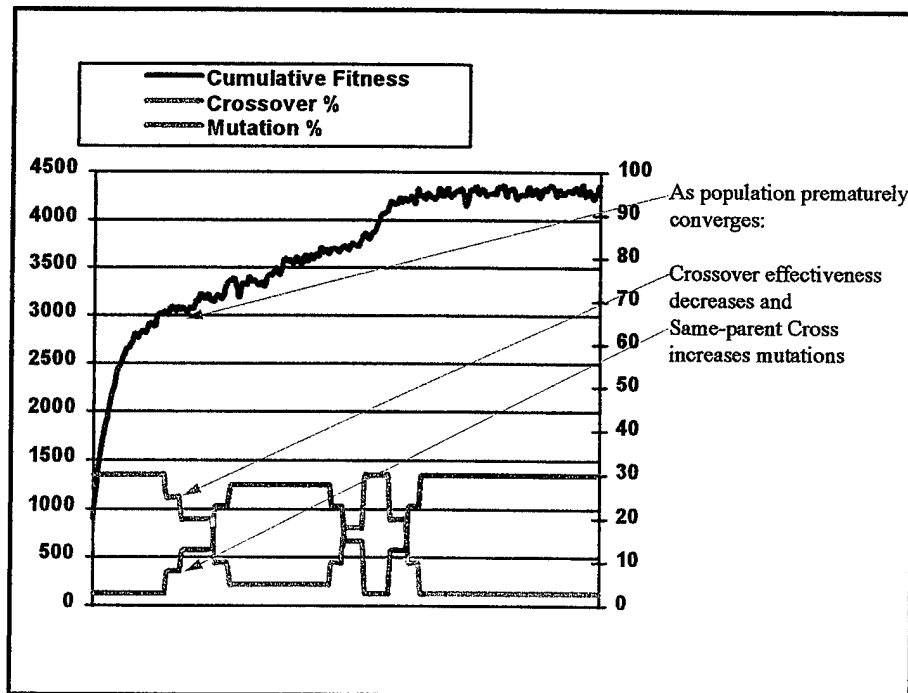


Figure 13. Effect of Same-parent Crossover Randomization

Finally, it was noted that since a genetic algorithm is based on probabilistic selection, some extremely strong rules failed to survive (by sheer chance) despite their selective advantage. This is an understandable consequence of natural selection; sometimes more capable species die solely because of "bad luck." The author reserved several spaces on the roulette wheel for the rules with the highest fitness measure in the population, regardless of their selection by the algorithm. This ensures that an extremely "good" rule will continue to be available for selection and recombination in successive generations.

2. Processing Speed Issues

However sophisticated the search technique may be, we must still keep the magnitude of this search problem in mind. One of our research goals was to ensure that the technology created did not require sophisticated, expensive, or proprietary hardware or software. For this reason the DaMI application was developed to run on a 80486/66Mhz personal computer using the Microsoft Window 3.xx or Windows 95 operating system. (Pentium 166's are used for production runs.) A very simple problem such as analyzing relations between 15 standard symptoms and 21 diagnoses (Boolean fields) yields a search space of 69 billion combinations. A 486 computer, using the "brute force" method, can test about 600,000 hypotheses (rules) per day. At that rate, this problem would take more than 315 years to complete. Even if the speed of processing could be accelerated by a factor of 100, the problem would still be impractically large. We have processed runs involving exposures/demographics and diagnoses that were on the order of $9.457 * 10^{16}$. Actual processing benchmarks are included later in the paper, but the point for the moment is that results using genetic algorithms take days not minutes to achieve.

Naturally the author took several steps to enhance speed on the given PC architecture. First, the population of rules is maintained in a RAM-based array space as is the statistical package's attribute and possible value matrix. This allows the genetic operations to be carried out with extreme speed. Task complexity is not really a speed issue at all for the genetic algorithm package; unfortunately, the database under analysis cannot be placed in RAM, so the statistical package becomes the speed limiting operation. Genetic operations take several seconds per

population, but the statistical package may take hours to analyze a single, large population. In the case of the statistical package, number of attribute and possible values is much more significant than the number of records in the analysis database. If the operating architecture could be enhanced to allow the genetic algorithm to pass statistical requests to multiple personal computer nodes, a significant processing advantage could be attained.

The nature of our research question concerning a possible syndrome affecting Persian Gulf War participants limits the complexity requirement of rules generated. In other words, rules involving too many attributes may be statistically significant, but are so specific that they may only describe a single participant. Naturally, these rules may have a selective advantage over less specific rules, because a single outlier reporting a highly unusual combination of attributes will be very highly rated. However, rules involving a single individual do not suggest a syndrome, which by definition is a series of conditions affecting a *group* of individuals. Therefore, we included a tunable parameter which limits the maximum complexity of rules generated. Rules involving too many attributes are given a low fitness function and are not sent to the statistical analysis package. It should be obvious that increasing the number of attributes in a single rule exponentially increases the complexity of the analysis by the search package.

3. Tuning the Fitness Measure, Verification, and Validation

One of greatest challenges faced is to develop a fitness that accurately reflects the requirements of CCEP medical researchers. It is critical that feedback is obtained at every step of the discovery process.

EXAMPLE *Just because there is a high association between hair loss and chronic fatigue syndrome within the database under examination does not mean that this is of any medical significance.*

It must also be understood that our technique has drastically reduced the number of correlations to be investigated by medical researchers, but it does not guarantee that each rule is

of value. That knowledge can only be obtained from medical professionals. Our goal is to provide a catalyst for their research and a "jumping off point" for more in-depth clinical investigation. If that mindset is maintained, the genetic algorithm is proving most helpful.

Verification is also a key issue. Rules and their associated fitness measures generated by a genetic algorithm will be true. That has been easily verified by conventional query. Ensuring that the rules generated are the best ones to describe the analysis database is more challenging. We have two different methods for responding to this challenge, duplicability, and reproducibility.

The database of 19,000 records has been split into several sample sets. Each sample set is selected randomly without replacement. We actually use two database subsets of around 7,700 records each. The genetic algorithm is applied to one sample subset and its output rules are then applied to the second subset. If the fitness measure for a rule is uniform throughout the two independent, randomly-selected databases, then there is confidence that this rule holds for the entire database and is not a statistical anomaly. We call this attribute *duplicability*.

The second verification procedure is *reproducibility*. It cannot be proven that a genetic algorithm has actually found the best rules for a given search space. The only way to accomplish this is to actually check every possible combination, which we have already stated is physically impractical. How then may we have any certainty that the technique has worked; that the algorithm has used a sufficiently large population over a sufficiently large number of generations to achieve an acceptable answer? Since a genetic algorithm depends on the simulation of survival of the fittest (Darwinism) based solely on probability modeling and random number generation, it will never analyze the same problem the same way twice. We run every problem twice and note the number of rules that occur in both outcome rule sets. If both independent discovery sessions produce a high number of the rule intersections, then this indicates that the state space has been searched exhaustively (see Figures #14 and #15). If this is not the case, then the population size and/or number of generations must be increased for an effective discovery session.

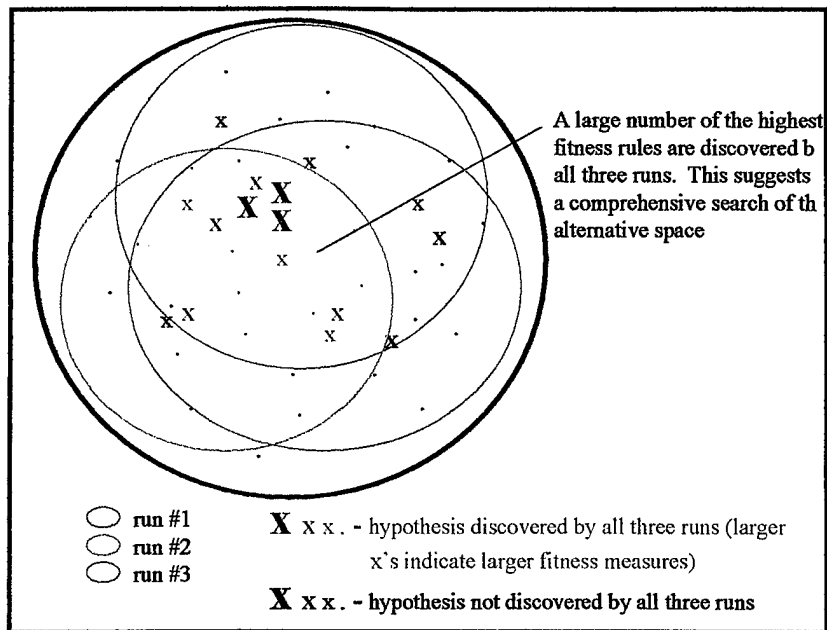


Figure 14. Strong Reproducibility in GA Search

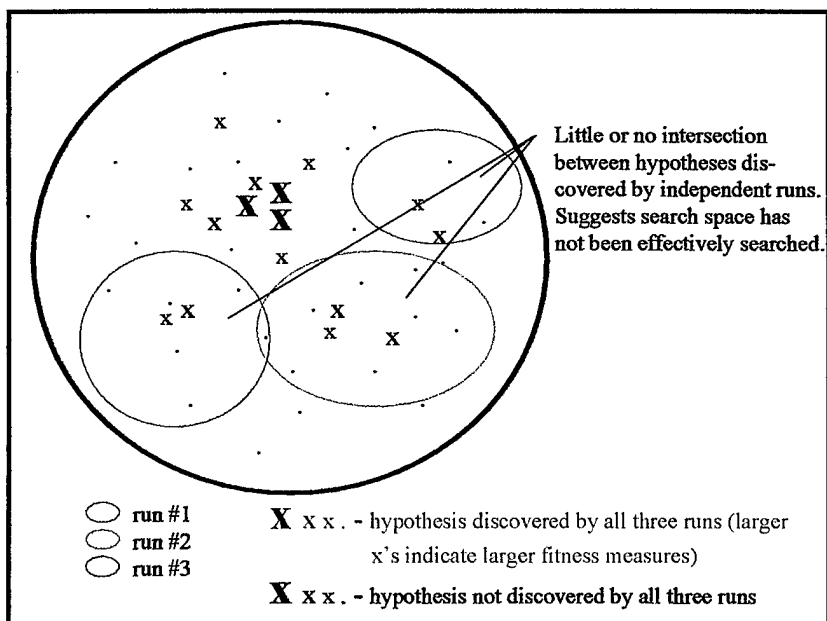


Figure 15. Weak Reproducibility of GA Search

Finally, a great deal of emphasis is placed on the discovery of rules which are intuitively obvious to medical professionals. This may appear insignificant at first, but as mentioned before

genetic algorithms are unguided random processes possessing *no knowledge of medical facts*. If, through their learning process, they produce a series of rules that mimic accepted medical knowledge then this lends confidence that accompanying rules, which do not make intuitive sense, may contain new and significant information.

[THIS PAGE INTENTIONALLY LEFT BLANK]

VI. RESULTS

A. SUMMARY

DaMI has achieved striking successes throughout our experiments. The theoretical basis for the design of this search algorithm is sound and has allowed this system to perform and produce results. DaMI is a very exciting application because its performance matches or exceeds theoretical expectations, and it identifies previously undiscovered correlations in the CCEP Desert Storm Database. In this chapter, we will characterize the initial success of DaMI by presenting a series of experimental results which build on the framework developed by this thesis. *Success* in this research is metered by responding to the following questions:

- Did the Genetic Algorithm (DaMI) perform as theoretically predicted?
- What correlations did the Genetic Algorithm actually find in the CCEP database, and were these hypotheses, at least from a statistical perspective, consistent with the research goals?
- How useful were the hypotheses discovered to CCEP medical researchers?

Each will be examined individually in the following sections of this chapter, building up to a comprehensive evaluation of DaMI's theoretical as well as practical performance.

Twenty-five discovery sessions (runs) have been conducted by DaMI thus far, of which six production runs are discussed in the results section. Earlier runs were used to test the performance of DaMI during development and refine the settings of tunable parameters for optimal discovery. Genetic algorithm development is a constant process of discovery, feedback and refinement. The runs conducted to date are by no means all-inclusive, but rather chronicle a successful venture into the CCEP database.

DaMI has been directed to analyze two different perspectives of the CCEP database (three identical production runs for each perspective). The first runs search for associations between the gender, service, race, and reported exposures of PGW participants (LHS) and the

diagnoses that were assigned by the CCEP medical examination process (RHS). We refer to these runs as *exposure-to-diagnosis* runs. The second set of runs search for associations between gender, service, race, and reported exposures of PGW participants (LHS) and the standard symptoms that were elicited during the CCEP medical examinations (RHS). We refer to these runs as *exposure-to-symptom* runs. The reader is referred to Appendix A for a detailed list of fields included in each analysis. Each production run utilized a population size of 1000, cross-over probability of 30%, mutation probability of 3.0% (see section V.C for a discussion of tunable parameters). Modified j-measure has been used as a fitness measure, and only the single best j-measure of all combinations of each individual attribute set was used for aggregate fitness by the statistical analysis package (see section V.B). Hypotheses generated were limited to combinations of up to three LHS attributes and two RHS attributes. Production runs have simulated at least 130 generations; some were allowed to continue for 170 generations.

B. DID THE GENETIC ALGORITHM PERFORM AS EXPECTED?

As theoretically predicted, DaMI performs very well, in terms of speed, hypothesis quality improvement, and search space coverage. This question focuses solely on the ability of DaMI to perform an efficient, self-improving search and not on the value of results to medical professionals (which will be discussed in the next section). The tremendous size of the search space has been mentioned earlier, but the number of possible combinations should be presented specifically at this point:

- **Exposure-to-diagnosis Runs.** 29 Boolean reported exposures, gender (2 possible values), service (6 values), race (8 values), and 21 Boolean diagnoses.

$$\text{Possible combinations} = 2^{29} \times 2 \times 6 \times 8 \times 2^{21} = 9.46 \times 10^{16}$$

- **Exposure-to-symptom Runs.** 29 Boolean reported exposures, gender (2 possible values), service (6 values), race (8 values), and 21 Boolean symptoms.

$$\text{Possible combinations} = 2^{29} \times 2 \times 6 \times 7 \times 2^{15} = 1.48 \times 10^{15}$$

It is clear that these two types of runs present a credible challenge to any genetic algorithm. They are both computationally explosive (because of search space size) and highly unstructured (because of the high number of LHS and especially RHS attributes), yet DaMI has processed them with striking success.

1. Analysis Speed

DaMI's search efficiency allows it to perform analyses, which normally take years, in a matter of hours. Analysis speed is the time required for a genetic algorithm to comprehensively search the given space. Comprehensive search will be dealt with shortly, but at the moment, we will focus on the time required for DaMI to complete an analysis. If that time is significantly less than would be possible using a "brute force" examination of the same database, then the first advantage has been achieved. As mentioned in section II, it was observed that a personal computer can test about 600,000 possible combinations per day. If that is the case, then the exposure to diagnosis run should take about 432 billion years—this is clearly not acceptable. Since DaMI never searches a space the same way twice, analysis times for the same problem vary; however, DaMI performs the same analysis in 36 hours (on average). Exposure-to-symptom runs take about 44 hours, using the genetic algorithm. Although the exposure-to-symptom runs involve a smaller search space, DaMI requires more generations to converge on an answer. Analysis times do increase in relation to the number of possible combinations; however, the character of the research question also affects the time required for DaMI to converge on an answer. Analysis times of similar runs are fairly consistent (less than 10% deviation). A profile of the three DaMI exposure-to-diagnosis runs is illustrated in Figure #16.

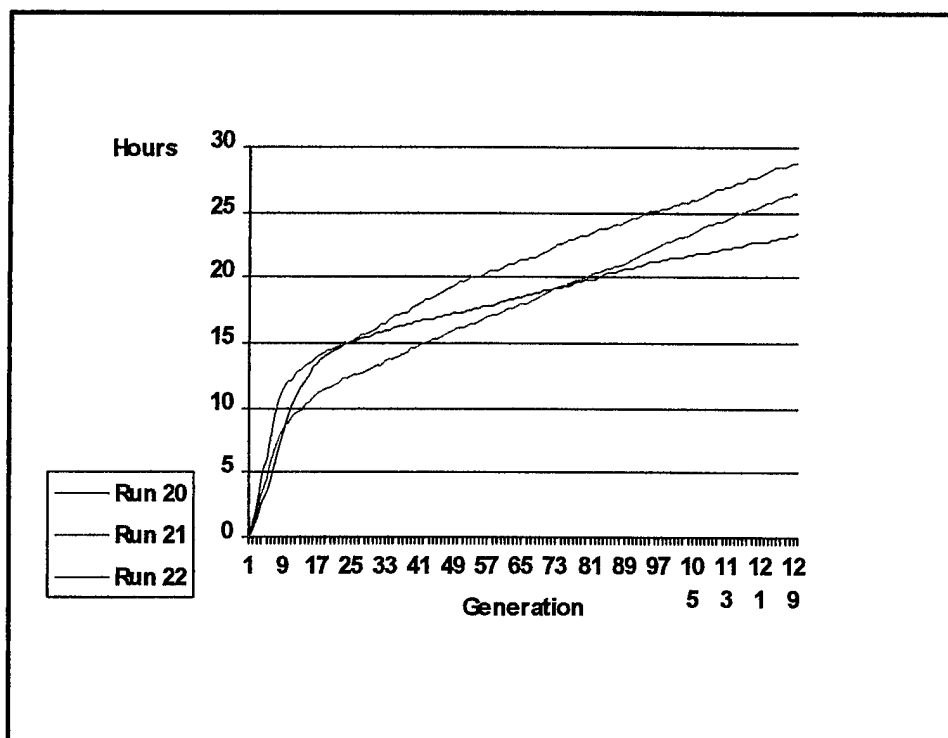


Figure 16. Analysis Speed Profile of Exposure-to-diagnosis Runs

Notice that the processing speed increases as a small group of rules begin to dominate the population (convergence). It must be reiterated that DaMI uses the same platform as was used for “brute force” testing; it is the selectivity of search (knowing what alternatives need not be tested) that gives this methodology its incredible advantage.

2. Hypothesis Quality Improvement

DaMI is consistently able to adaptively improve the quality of the hypotheses it generates as the analysis progresses. A genetic algorithm is theoretically an intelligent, adaptive search technique. This means that as processing time passes, the system will generate hypotheses of increasing quality based on the results of analyses already conducted. In the case of DaMI, this means quality is indicated by the fitness measure of a hypothesis. The cumulative fitness of a generation represents the aggregate quality of all the hypotheses synthesized during that generation. Although some new individuals in each generation may receive very low fitness

measures, if the cumulative fitness increases in successive generations, then the quality of hypotheses as a whole are improving. DaMI demonstrates the characteristic ability of genetic algorithms to rapidly increase the quality of new hypotheses generated. DaMI rapidly improves cumulative fitness until a small group of rules begins to dominate the population [premature convergence (Koza, 1989)], but (largely because of same-parent crossover randomization) it then boosts mutation probability and continues to break through to higher cumulative fitness plateaus. A profile of improving hypothesis quality for exposure-to-diagnosis runs is presented in Figure #17. Note that in each of the three runs, the cumulative fitness curve levels (signaling premature convergence) and then continues to sporadically increase.

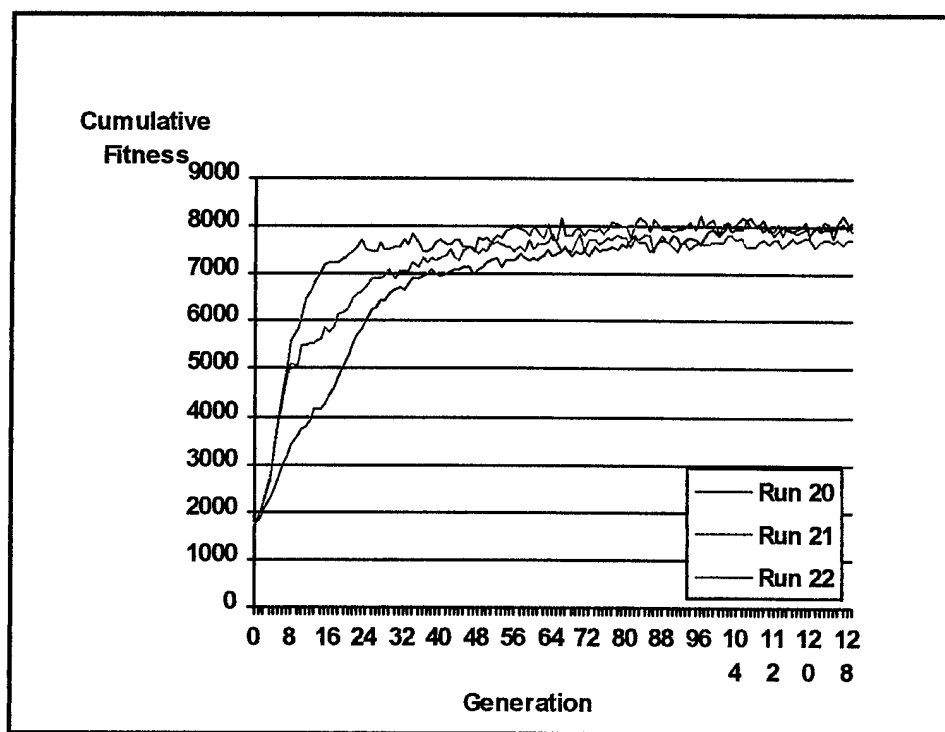


Figure 17. Analysis Speed Profile of Exposure to Diagnosis Runs

3. Reproducibility: Search Space Coverage

While a genetic algorithm may complete a search quickly, the speed advantage is of limited value without some indication that the results derived are actually the best in the search

space. DaMI produces consistent reproducibility on the extremely large spaces it searches, attesting to its strong ability to search a large space by testing a small subset of possible combinations. As discussed in section V.D.3, *proving* that a genetic algorithm has completely examined a space is a paradoxical question—you cannot prove that the genetic algorithm made the right decision without testing every possible hypothesis. Reproducibility gives a strong indication that the alternative space has been searched effectively. Ideally, we would like multiple independent runs of the genetic algorithm (see section V.D.3) in order to test only a few of the same rules of low fitness but converge on the same rules of high fitness. A low intersection of low fitness rules between runs indicates that each approached convergence from different areas of the search space (i.e. they did not all follow the same path). A high intersection of high fitness rules suggests that, despite entering the search space from different *directions*, each independent run has arrived at the same answer. This reproducibility strongly suggests that the entire search space has been effectively, but not physically, examined.

DaMI achieves high reproducibility in spite of the rapid search time and tremendous space. In the exposure-to-diagnosis study, all three runs agree on the same 16 highest fitness hypotheses. Lower fitness hypotheses show steadily decreasing levels of intersection, as is theoretically predicted. This is particularly exciting, because each production run has achieved consensus by testing only 7,100 - 7,400 of the 1,041,000 possible attribute combinations. The probability of three independent runs randomly agreeing on the same sixteen hypotheses (especially since each run is testing only 0.7 % of all possible attribute combinations) is infinitesimally small. The natural question is, "Did the three runs, by some streak of luck, enter the search space from the same starting point?" This is not the case, because the three runs only tested 14.1% of the same lower fitness rules, proving that they have entered the space from different points but converged on the same answer. Note in Figure #18 that the percentage of rule intersection (Runs 20, 21, and 22 are the three runs conducted in the exposure-to-diagnosis study) between runs approaches 100% for rules with a fitness measure higher than 8.0. This intersection decreases steadily as the fitness measure decreases (going left on the graph). In the case of exposure-to-symptoms, the reproducibility is not as high, but still quite striking. In this study, each run tested between 8,000 and 10,000 hypotheses. The three runs agree on 5 of 6 highest fitness hypotheses. This is represented in Figure #19 by an intersection percentage of

80% on hypotheses with a fitness of over 5.31 (Runs 23, 24, and 25 are the three runs conducted in the exposure-to-symptom study). Notice that, as in the exposure-to-diagnosis study, the intersection between runs decreases as the fitness measure decreases, culminating with an intersection of only 20% for rules with fitness measures between 1.0 and 3.0.

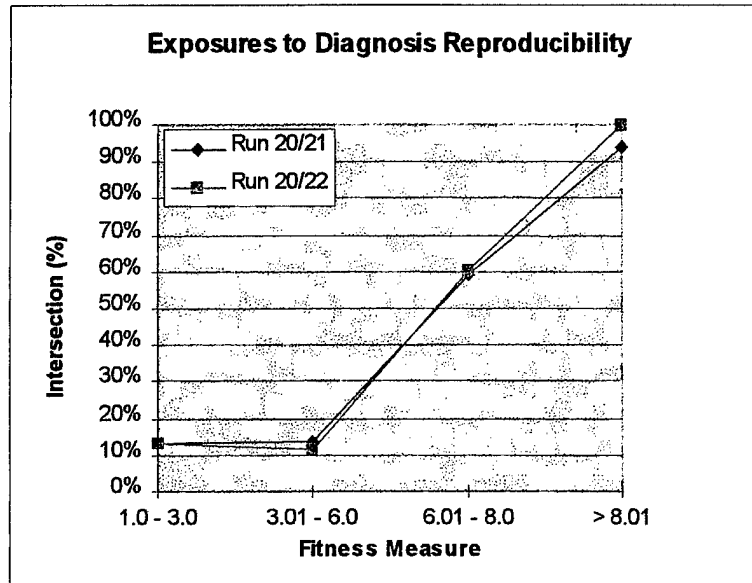


Figure 18. Exposure-to-diagnosis Reproducibility

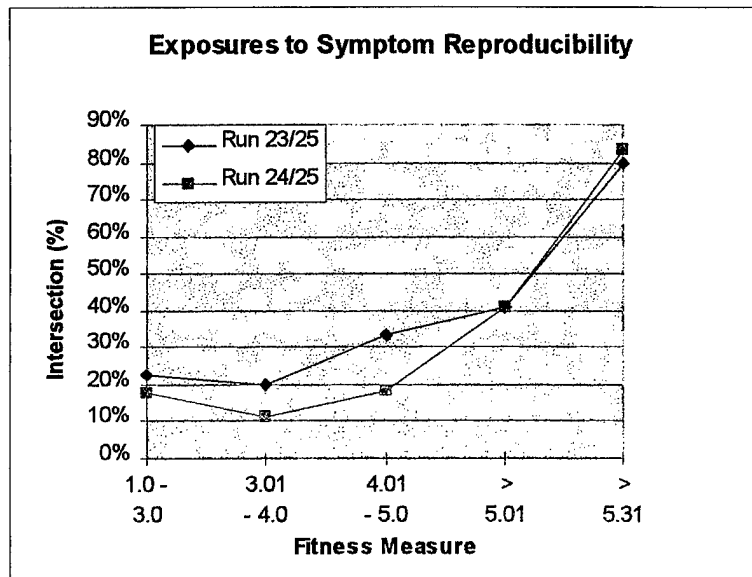


Figure 19. Exposure-to-symptom Reproducibility

Based on the high reproducibility of DaMI production runs, there is a strong indication that the search space has been effectively searched for the given fitness measure and search parameters. This is particularly significant in the case of Desert Storm research. Recall that the existence of any syndrome has not yet been determined. Therefore, if DaMI fails to find a viable syndrome profile but can show that the space has been searched effectively, that information will be of extremely high value to CCEP research. Additionally, any comprehensive list of correlations between risk factors and medical outcomes will be of value to PGW participants and the medical practitioners providing their ongoing medical care.

C. WHAT DID DaMI FIND?

DaMI has proven, by the standards of genetic algorithm theory, that it has studied the CCEP database quickly, intelligently, and comprehensively. All of the theory and development strategies now come down to one question, "What did we learn?" Computational results so far suggest that our system has succeeded at the given tasks, requiring relatively few resources. Experiments reveal no single syndrome, but numerous correlations do exist that require additional clinical analysis.

Based on DaMI research, there is no indication that a single syndrome or other medical entity is causing wide-spread adverse health ramifications among a significant cross-section of PGW participants in the CCEP program. By "significant," we mean that no group of over 100 participants, sharing a common reported exposure/demographic information, exhibit a unique set of reported symptoms and/or outcome diagnoses. Keep in mind that only the 21 most frequently reported diagnoses (and combinations of these) have been tested to date. This does not mean that a syndrome cannot exist, but the data collected by CCEP and specifically studied by this research does not indicate such a correlation.

There are, however, numerous correlations of exposure/demographic information and associated symptoms/diagnoses which suggest that smaller groups may share common health conditions based on shared exposure to common health risk factors. These associations are based solely on statistical correlation; therefore, a final determination is withheld pending review of the

information by medical professionals. In any case, the examined data suggests a need for further research.

The number of correlations found by DaMI is quite large; we have resisted summarizing hypotheses to preserve the robustness of the information. Therefore, the challenge of filtering and reporting awaits the input of CCEP researchers. Each exposure-to-diagnosis run has produced around 4,500 hypotheses, and each exposure-to-symptom run has produced about 6,100 hypotheses. In each case, the three sets of rules are combined into a single hypothesis set (with duplicates removed). The information has been further refined, subject to the following criteria:

- Hypotheses applying to fewer than five individuals in the sample set have been removed to prevent undue influence by single outliers. By definition, a syndrome is a medical condition shared by a number of individuals.
- Hypotheses are derived from a randomly selected 45% sample (without replacement) subset of the entire CCEP database. These hypotheses are tested against a separate 45% (independent) partition of the CCEP database. Hypotheses whose fitness measure in the second (verification) sample differed from the fitness measure from the original sample by more than 20% have been eliminated. Fitness measures which remain constant over both the original and verification sample are called *duplicable*, suggesting they hold true for the entire database and are not a statistical anomaly.

The application of the aforementioned selection criteria has resulted in a set of 2,653 candidate hypotheses concerning exposure-to-diagnoses and 4,959 hypotheses concerning exposure-to-symptoms. No minimum fitness measure threshold has been applied because the modified j-measure is an arbitrary score, suitable for ranking the order of interest of competing hypotheses. The fitness measure may not be attached to a specific interest "level." Obviously, a great number of the hypotheses having low fitness measures do not contain correlations strong enough to support strong research attention. For this reason and for the sake of brevity, only the 100 highest fitness hypotheses of each study are included in Appendix C and discussed in the next two result summary sections.

These two sections will discuss the highlights and some specific hypotheses from both the exposure-to-diagnosis and exposure-to-symptom studies. The exposure-to-diagnosis and exposure-to-symptom results are each exciting for different reasons. The exposure-to-diagnosis study contains many high confidence correlations--hypotheses which are applicable to over 50% of the participants concerned. The exposure-to-diagnosis hypotheses contain few unexpected correlations, but clearly demonstrate the ability of DaMI to cull out extremely strong correlations from a "mountain" of data. The exposure-to-symptom results contain many unexpected hypotheses, but with somewhat lower correlation strength. The exposure-to-symptom results attest to the sensitivity of DaMI analysis and contain *new* (previously undiscovered) information which should attract expanded clinical research.

1. Exposure-to-diagnosis Correlations

The exposure-to-diagnosis study yields a large number of strong correlations (positive predictive values between exposure and diagnosis of over 50%) and provides corroboration to some intuitive aspects of medical relationships. Several new relationships have been identified, but few hold information that is unexpected by the non-medical analyst, at least when studied separately from associated symptoms. DaMI demonstrates a powerful ability to cull strong correlations from a large body of data, and in that respect, the results are very exciting. It must be reiterated that only combinations of the 21 most frequently occurring diagnoses have been considered at this point. However, a restructuring of the CCEP diagnosis representation which groups like diagnoses (with differing ICD codes) may bear even more information.

No single exposure or group of exposures appear(s) to dominate the resulting hypotheses set, unlike what will be seen in the exposure-to-symptom study. Several exposures (but no demographic attributes) appeared in many of the 100 highest fitness hypothesis. 19% of the hypotheses included participants who were wounded and another 19% included participants who saw casualties. Yet another 19% of hypotheses included participants who reported exposure to "other paints" and 12% reported exposures to nerve gas. At first, the fact that many hypotheses include wounded participants appears interesting because only 1% of participants in the CCEP

database have been wounded. Also, only 4% of CCEP participants report exposure to nerve gas, so that too seems to be highly represented in the hypotheses. Casualties and other points in hypotheses are less surprising since both have been highly reported by CCEP participants (50% and 38% respectively). However, 37% of the hypotheses discovered include Post-traumatic Stress Disorder and 22% include Depression (CCEP, 1996, p 19). This high number of Psycho-social diagnosis prevalence in the hypothesis set decreases the surprise that many hypotheses concern wounded participants (as the two are commonly associated). Surprisingly, Severe Sleep Apnea is included in 20% of the hypotheses. Sleep Apnea is a medical condition not commonly linked to any CCEP reported exposure. This leaves only the prevalence of reported Nerve Gas exposures and the diagnosis of Sleep Apnea in hypotheses as the only unexpected attributes, from a macro perspective. Reported nerve gas exposure is all the more unexpected because chemical alarms and mustard gas (similar participant concerns) are notably scarce from the hypotheses. It will be seen later that reported nerve gas exposure plays a significant role in the exposure-to-symptom study. Finally, it should be noted that oil and smoke, heat and smoke, Pyridistine Hydrobromide (Pb), and headaches are included in few hypotheses—all are factors receiving high attention in CCEP research.

An explanation of the DaMI reporting format is included in Figure #20. While the space is not available to discuss even the 100 highest fitness hypotheses, several illustrative hypotheses are presented now in Figure #21. Especially in the exposures-to-diagnosis study, DaMI demonstrates the ability to unmask high level of association between exposure/demographic and diagnosis attributes. This association is not limited to high positive predictive value (high probability of *then* condition given the *if* condition), but is also able to look at the associations in reverse (high probability of *if* condition given the *then* condition) and examine the contraindications (*if* condition precludes the *then* condition) between exposures/demographics and diagnoses. An example of each association type is presented below. The medical professional is referred to Appendix C for a complete list of hypotheses.

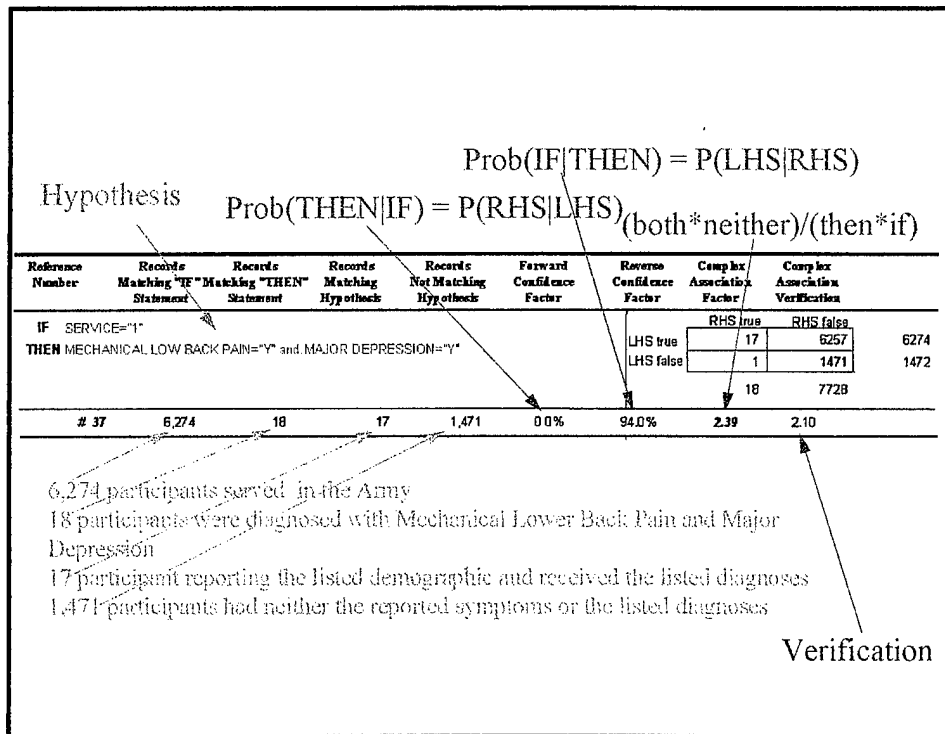


Figure 20. How to Read a DaMI Report

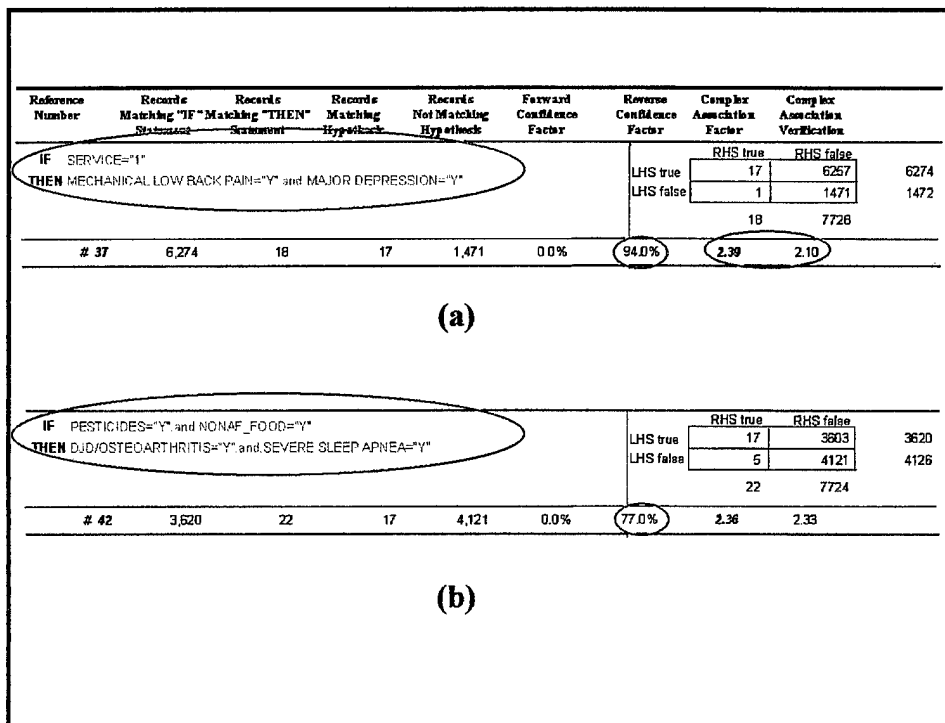


Figure 21. Exposure-to-diagnosis Examples

As stated before, the exposure-to-diagnosis examples presented here demonstrate the capability of DaMI to dig into a “mountain” of data and find strong hypotheses. The examples selected for presentation here are selected to illustrate that capability. It is highly recommended that the medical professional examine all of the hypotheses (Appendix C) in detail. Figure #21(a) is a hypothesis of extremely high positive predictive value. The hypothesis states that 94% of participants diagnosed with mechanical lower back pain and major depression served in the Army. 94% is an extremely high correlation for such a broad hypothesis (a specific diagnosis combination is linked to a single service). Note that both the fitness measure obtained using the analysis database (*complex association factor*) is quite close (2.39/2.10) to that of the verification database (*complex association verification*), suggesting that the rule holds for all participants (not a statistical anomaly). The hypothesis illustrated in figure #21(b) is much more specific, but is still quite strong. This hypothesis states that 77% of the participants diagnosed with DJD/Osteoarthritis and Severe Sleep Apnea reported eating Non-allied Forces food and reported exposure to pesticides. DaMI is capable of isolating strong data correlations, regardless of hypotheses specificity.

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification												
IF SERVICE="4" and PESTICIDES="Y" and MALARIA="Y" THEN ASTHMA="Y"							<table border="1"> <tr> <td></td> <td>RHS true</td> <td>RHS false</td> </tr> <tr> <td>LHS true</td> <td>9</td> <td>49</td> </tr> <tr> <td>LHS false</td> <td>379</td> <td>7318</td> </tr> <tr> <td></td> <td>388</td> <td>7358</td> </tr> </table>		RHS true	RHS false	LHS true	9	49	LHS false	379	7318		388	7358	49 7697
	RHS true	RHS false																		
LHS true	9	49																		
LHS false	379	7318																		
	388	7358																		
# 27	49	388	9	7,318	18.0%	2.0%	2.47	2.06												

(a)

IF PYRIDOSTIG="N" and CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y"							<table border="1"> <tr> <td></td> <td>RHS true</td> <td>RHS false</td> </tr> <tr> <td>LHS true</td> <td>9</td> <td>598</td> </tr> <tr> <td>LHS false</td> <td>424</td> <td>6724</td> </tr> <tr> <td></td> <td>433</td> <td>7313</td> </tr> </table>		RHS true	RHS false	LHS true	9	598	LHS false	424	6724		433	7313	598 7148
	RHS true	RHS false																		
LHS true	9	598																		
LHS false	424	6724																		
	433	7313																		
# 30	598	433	9	6,724	2.0%	2.0%	2.42	2.37												

(b)

Figure 22. Exposure-to-diagnosis Examples

The next two hypotheses are equally interesting, but are much more difficult to find using conventional search techniques. DaMI, using the Modified J-measure is able to see correlations which do not fit the high positive predictive value paradigm. The hypothesis in Figure #22(a) states that 18% of Marine participants reporting exposure to pesticides and malaria have been diagnosed with asthma. A positive predictive value of 18% does not jump out at the analyst and would therefore not figure prominently in a conventional analysis; however, DaMI notes that only 5.1% of all participants have been diagnosed with Asthma. This means that Marines reporting pesticide and malaria exposure are 3.5 times more likely to have been diagnosed with Asthma than the general CCEP participant population. In light of that fact, the 18% positive predictive value of this hypothesis is indeed significant, and DaMI has assigned it a high fitness measure. The hypothesis in Figure #22(b) is an example of *contraindication*. Note that this hypothesis shows no high correlation in either direction. The hypothesis states that 2% of participants reporting no exposure to Pb and not viewing casualties have been diagnosed with Post-traumatic Stress Disorder (PTSD). The reader's attention is directed to the matrix on the

right section of the hypothesis report. In 589 cases where the LHS is true, the RHS is false. Also, in 424 cases where the RHS is true, the LHS is false. 1,022 participants report information that in some way involves this hypothesis' exposures or diagnosis. In 99% of those cases, the exposures exclude the diagnosis outcome. In plain English, not reporting exposure to Pb or casualties precludes a diagnosis of PTSD. This fact, although readily apparent to conventional analysis, is very informative because of its exclusive properties and is therefore flagged by DaMI.

The exposure-to-diagnosis study hypotheses exemplify the ability of our genetic algorithm to find both strong, obvious correlations and more intricate associations in the CCEP database. Many of the hypotheses reinforce "common sense" medical knowledge, but remember that DaMI has discovered these hypotheses without the benefit of prior medical knowledge of any kind. In light of this success, serious attention should be directed toward those hypotheses presented that do not conform to present-day medical perceptions.

2. Exposure-to-symptom Correlations

The exposure-to-symptom study is more comprehensive than the diagnosis studies because the exposure-to-symptom runs consider every reported symptom category, not a top stratification. Many individual hypotheses contain *new* (or unexpected) correlations and there also several interesting trends revealed the about hypotheses as a group. This previously undiscovered information is of key interest to medical researchers. The author believes that this is the reason that exposure-to-symptom runs consistently take longer to converge and are somewhat less successful at reproducing than exposure-to-diagnosis runs. Even though the theoretical search space of exposure-to-symptom runs is smaller, the actual search space contains more represented combinations (because all attributes are included) and is therefore practically more difficult to solve. This explains the difference in run times for different studies noted previously.

While the exposure-to-diagnosis runs contain several intuitively obvious correlations, the exposure-to-symptom runs produce several strong but "unexpected" trends. These *unexpected* trends take the form of pervasive exposure and symptom combinations appearing in many of the

highest fitness hypotheses, despite the fact that these combinations are not prevalent in the CCEP database as a whole. These are the specific "threads" of information that DaMI has been designed to discover.

Several exposure attributes appear many times in the highest fitness exposure-to-symptom hypotheses:

- over 50% of the hypotheses include reported exposure to mustard gas (singly or in combination)
- almost 25% include reported exposure to nerve gas
- 14% include participants that were wounded in combat
- 12% include participants reporting some form of pre-conflict reproductive difficulties.

This is somewhat unusual because all of these attributes are reported relatively infrequently in the CCEP database as a whole. Mustard gas exposure has been reported by 2% of CCEP participants, nerve gas 6%, wounded in combat 2%, and pre-conflict reproductive difficulties 5.5% (CCEP, 1996, p. 19). Finally, the combination of reported nerve gas exposure and pre-conflict reproductive difficulties occurs in 9% of the top hypotheses. Notably scarce are hypotheses involving actual combat, chemical alarms, scud attacks, race, service, or post-conflict reproductive difficulties. It is surprising that since pre- and post-conflict reproductive difficulties are so highly statistically correlated, that post-conflict reproductive difficulties do not appear in any of the top hypotheses.

Similarly, the symptoms bleeding gums and weight loss are each included in over 50% of the hypotheses, and 44% of the hypotheses involve a combination of both bleeding gums and weight loss. Only 127 (or 1.6%) of the participants in the CCEP database subset studied (7746 total participants) reported that specific combination of symptoms. It is extremely interesting that so many hypotheses involve bleeding gums and weight loss, when these two symptoms are so scarce in the CCEP database at large. Also noteworthy is the large number of hypotheses relating reported mustard gas exposure to bleeding gums and weight loss (44% of hypotheses) and nerve gas exposure and pre-conflict reproductive difficulties with bleeding gums (9% of

hypotheses). Notably scarce in the hypotheses are hypotheses including joint pain, head aches, and fatigue, the symptoms most commonly elicited by physicians (CCEP, 1996, p. 20).

While thesis constraints prohibit discussing all 100 of the highest fitness hypotheses, several are included to illustrate some of the correlations discovered (Figure # 23).

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification
(a)								
IF SERVICE="Y" and DIESEL_FUEL="Y" and MUSTRD_GAS="Y" THEN DIFF="Y"							RHS true	RHS false
							LHS true	7 3
							LHS false	2143 5593
								2150 5596
# 25	10	2,150	7	5,593	70.0%	0.0%	2.81	2.34
(b)								
IF NERVE_GAS="Y" and PQ_PRIOR="Y" THEN BLEED="Y" and MUSCL="Y"							RHS true	RHS false
							LHS true	8 31
							LHS false	300 7407
								308 7438
# 19	39	308	8	7,407	21.0%	3.0%	2.85	2.43
(c)								
IF NERVE_GAS="Y" and MUSTRD_GAS="Y" THEN BLEED="Y" and WEIGH="Y"							RHS true	RHS false
							LHS true	7 75
							LHS false	120 7544
								127 7619
# 28	82	127	7	7,544	9.0%	6.0%	2.77	2.41

Figure #23. Exposures to Symptom Examples

The hypothesis in Figure #23(a) is included to demonstrate that DaMI, without the aid of medical knowledge, will discover intuitively obvious (to medical researcher) correlations. This hypothesis states that 70% of Navy participants who report exposure to diesel fuel and mustard gas also complain of difficulty breathing. It is understandable that anyone perceiving an exposure to mustard gas and who works with diesel fuel may, at some time, have suffered from difficulty breathing.

In Figure #23(b), it is noted that 21% of participants reporting exposure to nerve gas and pre-conflict reproductive difficulties complain of both bleeding gums and muscle pain. Note that the fitness measure (2.85) in the analysis database is very close to that of the verification

database (2.43), indicating that the hypothesis holds across different independent samples of the entire CCEP database. This hypothesis can be considered unexpected because this specific exposure combination is reported by only .5% of the participants and the symptomatology by only 3.9%.

In Figure #23(c), it is noted that 9% of participants reporting exposure to nerve gas and mustard gas, complain of both bleeding gums and weight loss. As before, the fitness measures (2.77/2.41) of both the analysis and verification database are quite close. Also note that this hypothesis holds in both directions; 6% of participants reporting bleeding gums and weight loss reported exposure to nerve gas and mustard gas. This hypothesis is also considered unexpected because this specific exposure combination is reported by only 1% of the participants and the symptomatology by only 1.6%.

In summation, the exposure-to-symptom study brings to light several correlations which warrant further clinical analysis. Interest lies, not only in the hypotheses themselves, but also in the high number of correlations involving rare combinations of exposures and symptoms.

D. ARE THE RESULTS USEFUL TO MEDICAL PROFESSIONALS?

The results of both the Exposure-to-diagnosis and Exposure-to-symptom studies and research methodology have been reviewed by Ph.D. Epidemiologists on the CCEP staff and the Director of the Deployment Surveillance Team. CCEP Epidemiologists feel that DaMI has great potential for "identifying previously unrecognized patterns of symptoms and diagnoses." (CCEP, Sep 1996) They also agree that DaMI has already identified many associations in the CCEP database that have not been found by conventional methods. However, they strongly emphasize that DaMI result hypotheses must be subjected to a more detailed, epidemiological-based post-processing before they can be of practical use to the CCEP research effort. They recommend that future DaMI research efforts be more closely coordinated with CCEP epidemiologists. The bottom line is that the substantial potential of DaMI as a research tool has been recognized by the medical researchers and the research sponsor has directed that DaMI be included actively in the study of Desert Storm Syndrome with the closer involvement of CCEP epidemiologists.

VII. CONCLUSION

After many months of theoretical development, genetic algorithm design, and fine tuning, DaMI has accomplished its goal—to comprehensively search the CCEP Desert Storm database and provide medical researchers with a subset of several thousand hypotheses for further investigation from the billions of possible combinations. DaMI has proven its ability to search an extremely large unstructured database and cull, in a reasonable amount of time, a subset of the highest interest rules within that database. DaMI has more to tell us about the CCEP database, as it can be retuned for different search priorities and measures of interest. It may also be applied to any number of similar bodies of medical and non-medical data.

This research began with a formidable analysis problem and an idea that the usefulness of computer analysis could extend beyond the conventional paradigm of “number crunching.” The author believed that by imparting a genetic algorithm with a model of a human researcher’s interest, that the genetic algorithm could intelligently attack a tremendous search problem and reduce it to a manageable size, given limited resources. We have taken a complex research question and unstructured database and formulated both into a workable representation of researcher interest and usable source of study. A genetic algorithm (DaMI) has been created which can perform a self-adapting, intelligent search with striking results. In short, DaMI has achieved our vision and exceeded our wildest expectations. This thesis has shown only one venture into this new realm of medical research, pre-emptive employment of genetic algorithm analysis; there are certainly many more adventures awaiting.

A. LESSONS LEARNED

The author encountered few problems during this thesis process. This thesis involves a very high visibility and politically sensitive subject, Desert Storm Syndrome. As such, there were numerous requirements for presentations and progress meetings in addition to the normal research challenges. Since the political obligations were linked to the feedback from the

sponsoring agency they could not be ignored; this placed a very high time demand on the author. Also, the sponsoring agency is located in Washington, D.C., so a great deal of travel and remote communication was required to ensure adequate project coordination. Finally, feedback for medical researchers in the field was very difficult to obtain because of their diverse geographic locations and limited availability.

The author has learned several valuable lessons from the thesis process:

- When doing a thesis involving data analysis, do not wait for results to start writing the thesis. A great deal of the thesis itself describes the theoretical basis and methodology of the research, and therefore, can be written before final results are achieved. The pressure of “doing the write-up” is a serious burden to good analysis and writing early helps to alleviate that pressure.
- If the thesis is directly funded by an outside agency (in my case the CCEP), it is important to clearly identify a liaison at that agency. In my case, there was not a clear procedure for information exchange established during the first half of the project, which made coordination haphazard. Once a clear coordination mechanism was put in place, the thesis process became much smoother.
- It is critical that a researcher have a sounding board who is not directly attached to the research. It was very easy for me to become so engrossed in the problem, that I began missing glaring solutions. I was lucky to have a single individual (not a genetic algorithm or medical expert per say) who reality checked my research and reviewed my thesis throughout my research. This feedback has proven invaluable to the quality of my thesis and the success of my research.

B. RECOMMENDATIONS FOR FUTURE RESEARCH

The success of DaMI opens the door to countless opportunities for future research. Two areas of study remain to be explored in the CCEP database:

- Analysis of demographic/exposure and a restructured diagnosis set. Efforts are currently underway to regroup participant diagnosis information so that similar diagnoses (even those with vastly divergent ICD codes) are grouped together. This will allow DaMI to analyze a majority of diagnoses, as opposed to the top 21 diagnoses as presented in this thesis.
- Analysis of time/motion study of units and their locations during the Persian Gulf Conflict. Since in many cases units are homogenous in location and therefore exposure to health risks, an analysis of the CCEP participants' unit location in time and associated symptoms and/or diagnoses should prove quite fruitful.

It should be obvious that DaMI has not been created with the sole intent of searching for a Desert Storm Syndrome. It is applicable to many other large, unstructured databases of medical and non-medical data. Aside from examining other bodies of data, there are several areas to investigate concerning DaMI itself:

- Comparison of DaMI performance with other commercial data mining software and other data mining techniques (like regression analysis, cluster analysis, and neural networks).
- Modification of DaMI's statistical package to use alternative fitness functions, such as Chi-square instead of just the Modified J-measure.
- Enhancement of the DaMI genetic algorithm to utilize parallel-processing for statistical computations. Clearly using a single PC is less efficient than a group of PC nodes operating simultaneously. This will dramatically increase search speed without increasing the complexity of computer hardware required.
- Rewriting of the DaMI code into C++ or Ada, so that it can run on a higher capacity computer platform. Of course, this will increase efficiency, but will make the algorithm more restrictive (less portable) in terms of operating platforms.

[THIS PAGE INTENTIONALLY LEFT BLANK]

APPENDIX A. CCEP DATA DICTIONARIES AND DATA COLLECTION METHODOLOGY

A. DATA DICTIONARY OF CCEP DATABASE

[THIS PAGE INTENTIONALLY LEFT BLANK]

CCEP DATA DICTIONARIES AND DATA COLLECTION METHODOLOGY						
	Def. Updatable:	Yes				
	Date Created:	10/5/95 3:21:36 PM				
	Last Updated:	10/5/95 3:35:06 PM				
	Record Count:	15467				
ID	Name	Data Type	Length	Usable	Problem	Action
1	PART_LNAME	Text	20	no	privacy act	Delete
2	PART_FNAME	Text	15	no	privacy act	Delete
3	PART_MNAME	Text	10	no	privacy act	Delete
4	PART_SSN	Text	11	no	privacy act	Delete
5	PAY_GRADE	Text	4	demographic		
6	SERVICE	Text	1	demographic		
7	REGION	Text	2	unk		
8	DMIS	Text	4	unk		
9	PART_BDAY	Date/Time	8	demographic		
10	PART_FMP	Text	2	demographic	change # to discrete	
11	SPON_SSN	Text	11	no	privacy act	Delete
12	SMOKE_NOW	Text	1	attribute	has U's	
13	NM_CG_NOW	Text	3	attribute ?		
14	SMOKE_PAST	Text	1	attribute	has U's	
15	NM_CG_PAST	Text	3	attribute ?		
16	OIL_SMOKE	Text	1	attribute	has U's	
17	HEAT_SMOKE	Text	1	attribute	has U's	
18	PASS_SMOKE	Text	1	attribute	has U's	
19	DIESL_FUEL	Text	1	attribute	has U's	
20	CARC_PAINT	Text	1	attribute	has U's	
21	OTHR_PAINT	Text	1	attribute	has U's	
22	OTHR_SOLVE	Text	1	attribute	has U's	
23	URANIUM	Text	1	attribute	has U's	
24	MICROWAVES	Text	1	attribute	has U's	
25	PESTICIDES	Text	1	attribute	has U's	
26	NERVE_GAS	Text	1	attribute	has U's	
27	PYRIDOSTIG	Text	1	attribute	has U's	
28	MUSTRD_GAS	Text	1	attribute	has U's	
29	CONTM_FOOD	Text	1	attribute	has U's	
30	CONTM_WATR	Text	1	attribute	has U's	
31	NONAF_WATR	Text	1	attribute	has U's	
32	NONAF_FOOD	Text	1	attribute	has U's	
33	ANTHRAX	Text	1	attribute	has U's	
34	BOTULISM	Text	1	attribute	has U's	
35	MALARIA	Text	1	attribute	has U's	
36	OTHER_EXP1	Text	35	attribute	has U's	
37	OTHER_EXP2	Text	35	attribute	has U's	
38	OTHER_EXP3	Text	35	attribute	has U's	
39	ACT_COMBAT	Text	1	attribute	has U's	

DATASTRU.XLS

40	WOUNDED	Text	1	attribute	has U's	
41	CASUALTIES	Text	1	attribute	has U's	
42	SCUD_ATTAC	Text	1	attribute	has U's	
43	CHEM_ALARM	Text	1	attribute	has U's	
44	PQ_CHD_P	Number (Dou	8	attribute		
45	PQ_CHD_A	Number (Dou	8	attribute		
46	PQ_INF_P	Text	1	attribute	combine into single field	
47	PQ_INF_A	Text	1	attribute	"	
48	PQ_MIS_P	Number (Dou	8	attribute	"	
49	PQ_MIS_A	Number (Dou	8	attribute	"	
50	PQ_SB_P	Number (Dou	8	attribute	"	
51	PQ_SB_A	Number (Dou	8	attribute	"	
52	PQ_ID_P	Number (Dou	8	attribute	"	
53	PQ_ID_A	Number (Dou	8	attribute	"	
54	PQ_DEF_P	Number (Dou	8	attribute	"	
55	PQ_DEF_A	Number (Dou	8	attribute	combine into single field	
56	SPON_LNAME	Text	20	no	privacy act	delete
57	SPON_FNAME	Text	11	no	privacy act	delete
58	SPON_MNAME	Text	11	no	privacy act	delete
59	SEX	Text	1	demographic	blanks	
60	RACE	Text	1	demographic	blanks	
61	MAR_STATUS	Text	1	demographic	blanks	
62	DUTY_STAT	Text	6	attribute	don't know code	
63	MOS_NEC_AF	Text	7	attribute	blanks (not too many)	
64	LOST_WORK	Number (Dou	8	maybe	question info value	LOFR
65	CHIEF_COMP	Text	35	no	text	delete
66	CHIEF_DTE	Date/Time	8	attribute ?	question info value	LOFR
67	CHIEF_DURA	Number (Dou	8	no	different for diff diags	delete
68	FATIG_DTE	Date/Time	8	maybe	question info value	LOFR
69	FATIG_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
70	ABDOM_DTE	Date/Time	8	maybe	question info value	LOFR
71	ABDOM_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
72	BLEED_DTE	Date/Time	8	maybe	question info value	LOFR
73	BLEED_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
74	DEPRE_DTE	Date/Time	8	maybe	question info value	LOFR
75	DEPRE_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
76	DIARR_DTE	Date/Time	8	maybe	question info value	LOFR
77	DIARR_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
78	DIFFI_DTE	Date/Time	8	maybe	question info value	LOFR
79	DIFFI_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
80	SHORT_DTE	Date/Time	8	maybe	question info value	LOFR
81	SHORT_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
82	HAIRL_DTE	Date/Time	8	maybe	question info value	LOFR
83	HAIRL_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
84	HEAD_A_DTE	Date/Time	8	maybe	question info value	LOFR
85	HEAD_A_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
86	JOINT_DTE	Date/Time	8	maybe	question info value	LOFR
87	JOINT_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
88	MEMOR_DTE	Date/Time	8	maybe	question info value	LOFR

DATASTRU.XLS

89	MEMOR_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
90	MUSCL_DTE	Date/Time	8	maybe	question info value	LOFR
91	MUSCL_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
92	RASH_DTE	Date/Time	8	maybe	question info value	LOFR
93	RASH_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
94	SLEEP_DTE	Date/Time	8	maybe	question info value	LOFR
95	SLEEP_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
96	WEIGH_DTE	Date/Time	8	maybe	question info value	LOFR
97	WEIGH_DURA	Number (Dou	8	attribute	number confuses algo	yes/no
98	OTHR1_COMP	Text	20	no	can't correlate text	delete
99	OTHR1_DTE	Date/Time	8	no	can't correlate text	delete
100	OTHR1_DURA	Number (Dou	8	no	can't correlate text	delete
101	OTHR2_COMP	Text	20	no	can't correlate text	delete
102	OTHR2_DTE	Date/Time	8	no	can't correlate text	delete
103	OTHR2_DURA	Number (Dou	8	no	can't correlate text	delete
104	OTHR3_COMP	Text	20	no	can't correlate text	delete
105	OTHR3_DTE	Date/Time	8	no	can't correlate text	delete
106	OTHR3_DURA	Number (Dou	8	no	can't correlate text	delete
107	OTHR4_COMP	Text	20	no	can't correlate text	delete
108	OTHR4_DTE	Date/Time	8	no	can't correlate text	delete
109	OTHR4_DURA	Number (Dou	8	no	can't correlate text	delete
110	PRI_DIAG	Text	40	no	text	delete
111	PRI_ICD	Text	6	RHS		
112	SEC_DIAG1	Text	40	no	text	delete
113	SEC_ICD1	Text	6	RHS	blanks	
114	SEC_DIAG2	Text	40	no	text	delete
115	SEC_ICD2	Text	6	RHS	blanks	
116	SEC_DIAG3	Text	40	no	text	delete
117	SEC_ICD3	Text	6	RHS	blanks	
118	SEC_DIAG4	Text	40	no	text	delete
119	SEC_ICD4	Text	6	RHS	blanks	
120	SEC_DIAG5	Text	40	no	text	delete
121	SEC_ICD5	Text	6	RHS	blanks	
122	SEC_DIAG6	Text	40	no	text	delete
123	SEC_ICD6	Text	6	RHS	blanks	
124	ALLER_CONS	Text	1	no	question info value	delete
125	AUDIO_CONS	Text	1	no	question info value	delete
126	CARDI_CONS	Text	1	no	question info value	delete
127	DENTL_CONS	Text	1	no	question info value	delete
128	DERMA_CONS	Text	1	no	question info value	delete
129	EARNT_CONS	Text	1	no	question info value	delete
130	ENDOC_CONS	Text	1	no	question info value	delete
131	GASTR_CONS	Text	1	no	question info value	delete
132	HEMAT_CONS	Text	1	no	question info value	delete
133	INFEC_CONS	Text	1	no	question info value	delete
134	NEPHR_CONS	Text	1	no	question info value	delete
135	NEURO_CONS	Text	1	no	question info value	delete
136	OCCUP_CONS	Text	1	no	question info value	delete
137	PULMO_CONS	Text	1	no	question info value	delete

DATASTRU.XLS

138	PSYCH_CONS	Text	1	no	question info value	delete
139	PTEST_CONS	Text	1	no	question info value	delete
140	RHEUM_CONS	Text	1	no	question info value	delete
141	MOVE_ON	Text	1	no	question info value	delete
142	DIAG_DTE	Date/Time	8	no	question info value	delete
143	DIAG_DONE	Text	1	no	question info value	delete
144	PTQS_DONE	Text	1	no	question info value	delete
145	PRQS_DONE	Text	1	no	question info value	delete
146	IREL_DONE	Text	1	no	question info value	delete
147	DECL_DONE	Text	1	no	question info value	delete
148	HOME_ADDR1	Text	30	no	privacy act	delete
149	HOME_ADDR2	Text	30	no	privacy act	delete
150	HOME_TOWN	Text	20	no	privacy act	delete
151	HOME_STATE	Text	2	demographic		
152	HOME_ZIP	Text	5	no	info too specific	delete
153	WORK_PHONE	Text	12	no	privacy act	delete
154	HOME_PHONE	Text	12	no	privacy act	delete
155	DCFORM_DTE	Date/Time	8	no	no info value	delete
156	STARTLATER	Text	1	no	no info value	delete
157	WHENTOCALL	Text	15	no	no info value	delete
158	DECLINE	Text	1	no	no info value	delete
159	WITHDRAW	Text	1	no	no info value	delete
160	EVAL_COMP	Text	1	no	no info value	delete
161	SATISFIED	Text	1	attribute ?	question info value	
162	PQ_DATE	Date/Time	8	no	no info value	delete
163	PQ_EVALDTE	Date/Time	8	no	no info value	delete
164	MIL_ADDR1	Text	30	no	no info value	delete
165	MIL_ADDR2	Text	30	no	no info value	delete
166	MIL_STATE	Text	2	no	no info value	delete
167	MIL_ZIP	Text	5	no	no info value	delete
168	CHECKL_DTE	Date/Time	8	no	no info value	delete
169	REPORT_DTE	Date/Time	8	no	no info value	delete
170	REPORT_TIM	Text	8	no	no info value	delete
171	PRIOR_JAN	Text	1	no	no info value	delete
172	REFUSED	Text	1	no	no info value	delete
173	NEGLECTED	Text	1	no	no info value	delete
174	EDS_VIEWED	Yes/No	1	no	no info value	delete
175	DCF_MISSIN	Text	1	no	no info value	delete
176	UIC	Text	8	attribute		
177	PHASE	Text	1	no	no info value	delete

B. DATA COLLECTION METHODS

This section is quoted directly from (CCEP, 1996, pp. 13-14)

Participants may enroll in the CCEP by calling a toll-free number (1-800-796-9699), which provides information and referrals to individuals requesting medical evaluations or by contacting their local military medical treatment facility (MTF). All MHSS eligible beneficiaries are eligible for the CCEP. For eligibility in the CCEP, a PGW veteran (or dependent) must have been eligible for DoD health care in June 1994 or later.

Once an individual is referred, the CCEP provides a two-phase, comprehensive medical evaluation, with Phase I being conducted at one of 184 local MTFs. Phase II (when required) is conducted at one of 14 regional medical centers (RMCs). The medical review includes questions about family history, health, occupation, and unique exposures in the Gulf War, as well as a structured review of symptoms.

Once a participant has completed the examination processes, copies of examination results are forwarded to the CCEP Program Management Team (PMT), where they undergo quality assurance procedures, and the data are entered into the master CCEP database.

Additionally, of those CCEP participants suffering chronic, debilitating symptoms, the DoD has established an SCC at Walter Reed Army Medical Center and will have a second center opening in mid 1996 at Wilford Hall Medical Center, Lackland AFT, Texas.

The data, which were initially entered into a relational database, were translated into a statistical format for this (*CCEP Report on 18,598 Participants*) report. Various validity checks were conducted to ensure that the data were appropriated for interpretation. Statistical tests and descriptive analyses were conducted on various categories of participants, including those in theater during the Persian Gulf War, their spouses, and their children. Moreover, the CCEP participants who were in theater were compared to the PGW population as a whole and were stratified by units to compare those units with higher CCEP participation to those units with lower CCEP participation. Specific analyses concerning self-reported exposures, physician-elicited symptoms, diagnoses, self-reported reproductive outcomes, self-reported lost workdays, physical evaluation boards (PEBs), and program satisfaction were conducted. Additionally, a

comparative analysis with the NAMCS data was conducted using age, sex, race, ethnicity, and diagnostic code variables to more closely match the CCEP population.

[THIS PAGE INTENTIONALLY LEFT BLANK]

APPENDIX B. DATA DICTIONARY OF SELECTED DaMI FILES

[THIS PAGE INTENTIONALLY LEFT BLANK]

Structure for table:
 Number of data records:
 Date of last update:
 Code Page:

C:\RESEARCH\VFP\VFPDOCS\DAMISAMP.DBF
 170340
 08/04/96
 1252

Field	Field Name Nulls	Type	Width	Dec	Index	Collate
1	RULE No	Integer	4			
2	CF No	Numeric	6	2		
3	CUMCF No	Numeric	6	2		
4	GENERATN No	Integer	4			
5	SERVICE No	Character	3			
6	SMOKE_NOW No	Character	3			
7	SMOKE_PAST No	Character	3			
8	OIL_SMOKE No	Character	3			
9	HEAT_SMOKE No	Character	3			
10	PASS_SMOKE No	Character	3			
11	DIESL_FUEL No	Character	3			
12	CARC_PAINT No	Character	3			
13	OTHR_PAINT No	Character	3			
14	OTHR_SOLVE No	Character	3			
15	URANIUM No	Character	3			
16	MICROWAVES No	Character	3			
17	PESTICIDES No	Character	3			
18	NERVE_GAS No	Character	3			
19	PYRIDOSTIG No	Character	3			
20	MUSTRD_GAS No	Character	3			
21	CONTM_FOOD No	Character	3			
22	CONTM_WATR No	Character	3			
23	NONAF_WATR No	Character	3			
24	NONAF_FOOD No	Character	3			
25	ANTHRAX No	Character	3			
26	BOTULISM No	Character	3			
27	MALARIA No	Character	3			
28	ACT_COMBAT No	Character	3			
29	WOUNDED No	Character	3			
30	CASUALTIES No	Character	3			
31	SCUD_ATTAC No	Character	3			
32	CHEM_ALARM No	Character	3			
33	PQ_PRIOR No	Character	3			
34	PQ_AFTER No	Character	3			

35	SEX No	Character	3
36	RACE No	Character	3
37	FATIG No	Character	3
38	ABDOM No	Character	3
39	BLEED No	Character	3
40	DEPRE No	Character	3
41	DIARR No	Character	3
42	DIFFI No	Character	3
43	SHORT No	Character	3
44	HAIRL No	Character	3
45	HEAD No	Character	3
46	JOINT No	Character	3
47	MEMOR No	Character	3
48	MUSCL No	Character	3
49	RASH No	Character	3
50	SLEEP No	Character	3
51	WEIGH No	Character	3

** Total **

162

Structure for table: C:\RESEARCH\VFP\VFPDOCS\RULELIB.DBF
 Number of data records: 5446
 Date of last update: 08/04/96
 Code Page: 1252

Field	Field Name	Type	Width	Dec	Index	Collate
	Nulls					
1	RULE_NUMBE	Numeric	8			
	No					
2	NO_TRUE_LH	Numeric	8			
	No					
3	NO_TRUE_RH	Numeric	8			
	No					
4	NO_TRUE_BO	Numeric	8			
	No					
5	NO_FALSE_B	Numeric	8			
	No					
6	STANDARD_C	Numeric	5	2		
	No					
7	REVERSE_CF	Numeric	5	2		
	No					
8	COMPLEX_CF	Numeric	5	2	Desc	Machi
	No					
9	VCOMPLEX	Numeric	5	2		
	No					
10	LHS_TEXT	Character	100			
	No					
11	RHS_TEXT	Character	100			
	No					
12	RHS_VERB	Character	150			
	No					
13	REF_NUM	Integer	4			
	No					
** Total **			415			

**APPENDIX C. TOP 100 HYPOTHESES DISCOVERED BY
EXPOSURES-TO-DIAGNOSIS AND EXPOSURE-TO-SYMPTOM
STUDIES**

[THIS PAGE INTENTIONALLY LEFT BLANK]



NPS Data Mining Initiative (DaMI)

09/06/96 Detailed Hypothesis Report: Exposure-to-diagnosis Study

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF WOUNDED="N"									
THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="Y"									
# 1	7,278	8	5	465	0.0%	63.0%	3.24	3.21	60.53006
IF CASUALTIES="Y"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.ASTHMA="Y"									
# 2	4,384	31	28	3,359	1.0%	90.0%	2.97	2.79	1.64458
IF SEX="F"									
THEN SEVERE SLEEP APNEA="Y"									
# 3	914	213	5	6,624	1.0%	2.0%	2.74	2.62	18.67753
IF PESTICIDES="N".and.CASUALTIES="Y".and.PQ_AFTER="Y"									
THEN CHRONIC FATIGUE="N".and.MAJOR DEPRESSION="N"									
# 4	93	7,298	92	447	99.0%	1.0%	2.74	2.67	556.80926
IF PYRIDOSTIG="N".and.CHEM_ALARM="N"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.DJD/OSTEOARTHRITIS="N"									
# 5	472	417	5	6,862	1.0%	1.0%	2.72	2.38	18.16343

Reference Number	Records Matching "If" Statement	Records Matching "Then" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SMOKE_NOW="N".and.DIET_FUEL="N".and.CONTM_FOOD="Y" THEN MECHANICAL LOW BACK PAIN="N".and.PERENNIAL ALLERGIC RHINITIS="N"									
# 6	36	6,681	35	1,064	97.0%	1.0%	2.72	2.54	76.47374
IF SEX="F" THEN MEMORY LOSS="N".and.SEVERE SLEEP APNEA="Y"									
# 7	914	204	5	6,633	1.0%	2.0%	2.70	2.59	17.47080
IF SEX="F" THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="N"									
# 8	914	205	5	6,632	1.0%	2.0%	2.70	2.57	17.60449
IF CHEM_ALARM="Y".and.RACE="C" THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="Y"									
# 9	2,849	8	6	4,895	0.0%	75.0%	2.64	2.47	0.03724
IF OTHR_PAINT="N".and.CONTM_FOOD="Y".and.CASUALTIES="N" THEN ASTHMA="N".and.SEVERE SLEEP APNEA="N"									
# 10	60	7,161	59	584	98.0%	1.0%	2.58	2.27	280.93742

Reference Number	Records Matching "IF" Matching "THEN" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 11	IF SERVICE="2" and.ACT_COMBAT="N" THEN POST-TRAUMATIC STRESS DISORDER="Y"								
	621	433	8	6,700	1.0%	2.0%	RHS true 8 425 433		
							RHS false 613 6700 7313		621 7125
							2.58	2.39	23.94975
# 12	IF PASS_SMOKE="Y" and.SEX="F" THEN SEVERE SLEEP APNEA="Y" and.MAJOR DEPRESSION="N"								
	810	205	5	6,736	1.0%	2.0%	RHS true 5 200 205		
							RHS false 805 6736 7541		810 6936
							2.56	2.46	14.28480
# 13	IF NERVE_GAS="Y" and.MUSTRD_GAS="Y" THEN CHRONIC FATIGUE="N" and.MAJOR DEPRESSION="Y"								
	82	180	8	7,492	10.0%	4.0%	RHS true 8 172 180		
							RHS false 74 7492 7566		82 7664
							2.55	2.43	13.46277
# 14	IF SEX="F" THEN CHRONIC MUSCLE TENSION HEADACHES="N" and.SEVERE SLEEP APNEA="Y"								
	914	176	5	6,661	1.0%	3.0%	RHS true 5 171 176		
							RHS false 909 6661 7570		914 6832
							2.54	2.51	13.77214
# 15	IF DIESEL_FUEL="Y" and.SEX="F" THEN SEVERE SLEEP APNEA="Y"								
	743	213	5	6,795	1.0%	2.0%	RHS true 5 208 213		
							RHS false 738 6795 7533		743 7003
							2.51	2.93	13.05346

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SERVICE="4".and.PESTICIDES="Y".and.MALARIA="Y" THEN ASTHMA="Y".and.SEVERE SLEEP APNEA="N"									
# 16	49	372	9	7,334	18.0%	2.0%	2.51	2.03	11.89928
IF OIL_SMOKE="Y".and.DIESL_FUEL="Y".and.CONTM_FOOD="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.HYPERTENSION="Y"									
# 17	1,410	18	9	6,327	1.0%	50.0%	2.51	2.20	3.16300
IF OTHR_PAINT="N".and.URANIUM="N".and.CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.HEADACHES, MIGRAINE="N"									
# 18	485	396	6	6,871	1.0%	2.0%	2.51	2.85	15.87867
IF SERVICE="3".and.NONAF_WATR="Y".and.CASUALTIES="Y" THEN MAJOR DEPRESSION="Y"									
# 19	64	183	6	7,505	9.0%	3.0%	2.48	2.53	8.73342
IF MUSTRD_GAS="Y".and.MALARIA="N" THEN IRRITABLE BOWEL SYNDROME="Y"									
# 20	23	467	5	7,261	22.0%	1.0%	2.47	2.18	4.90297

Reference Number	Records Matching "IF" Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square	
IF SERVICE="4".and.PESTICIDES="Y".and.MALARIA="Y" THEN ASTHMA="Y"						RHS true	RHS false		
					LHS true	9	40	49	
					LHS false	379	7318	7697	
						388	7358		
# 21	49	388	9	7,318	18.0%	2.0%	2.47	2.06	10.96573
IF OTHR_SOLVE="Y".and.NONAF_FOOD="Y".and.WOUNDED="N" THEN PERENNIAL ALLERGIC RHINITIS="Y".and.SEVERE SLEEP APNEA="Y"						RHS true	RHS false		
					LHS true	12	2451	2463	
					LHS false	6	5277	5283	
						18	7728		
# 22	2,463	18	12	5,277	0.0%	67.0%	2.46	2.17	0.58258
IF SMOKE_NOW="Y".and.DIESL_FUEL="N".and.CHEM_ALARM="Y" THEN IRRITABLE BOWEL SYNDROME="N".and.PAPULA ECZEMA="Y"						RHS true	RHS false		
					LHS true	7	38	45	
					LHS false	316	7385	7701	
						323	7423		
# 23	45	323	7	7,385	16.0%	2.0%	2.46	2.14	8.74542
IF CONTM_WATR="Y".and.WOUNDED="Y".and.CHEM_ALARM="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y"						RHS true	RHS false		
					LHS true	10	40	50	
					LHS false	423	7273	7696	
						433	7313		
# 24	50	433	10	7,273	20.0%	2.0%	2.46	2.24	11.53329
IF SMOKE_NOW="Y".and.PQ_PRIOR="N" THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="Y"						RHS true	RHS false		
					LHS true	5	2174	2179	
					LHS false	3	5564	5567	
						8	7738		
# 25	2,179	8	5	5,564	0.0%	63.0%	2.45	2.04	0.26834

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Verification	Chi-square
# 26	IF DIESEL_FUEL="Y".and.CONTM_FOOD="Y".and.WOUNDED="Y". THEN POST-TRAUMATIC STRESS DISORDER="Y"				20.0%	LHS true LHS false	RHS true RHS false	11 45 422 7268 433 7313	56 7690
					3.0%		2.44	2.17	12.43036
	IF HEAT_SMOKE="N".and.NONAF_WATR="N".and.PQ_PRIOR="Y". THEN CHRONIC FATIGUE="N".and.MAJOR DEPRESSION="N"					LHS true LHS false	RHS true RHS false	68 7230 7298 448	69 7677
# 27	IF URANIUM="N".and.CASUALTIES="N". THEN POST-TRAUMATIC STRESS DISORDER="Y"				99.0%	LHS true LHS false	RHS true RHS false	17 1072 416 6241 433 7313	1089 6657
					1.0%		2.44	2.43	430.30393
					4.0%		2.44	2.26	41.76943
# 28	IF CASUALTIES="N".and.CHEM_ALARM="N". THEN POST-TRAUMATIC STRESS DISORDER="Y".and.PAPULA ECZEMA="N"				2.0%	LHS true LHS false	RHS true RHS false	17 1116 396 6217 413 7333	1133 6613
					4.0%		2.43	2.15	41.45980
					2.0%		2.43	2.15	41.45980
# 29	IF PYRIDOSTIG="N".and.CASUALTIES="N". THEN POST-TRAUMATIC STRESS DISORDER="Y"					LHS true LHS false	RHS true RHS false	9 589 424 6724 433 7313	598 7148
					2.0%		2.42	2.37	20.89844
					2.0%		2.42	2.37	20.89844

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF OTHER_PAINT="Y".and.NONAF_FOOD="N".and.WOUNDED="Y" THEN POST-TRAUMATIC STRESS DISORDER="N".and.POLY ARTHRALGIAS="N"									
# 31	18	6,702	11	1,037	61.0%	0.0%	2.41	2.21	54.23568
IF DIESEL_FUEL="Y".and.CONTM_FOOD="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.HYPERTENSION="Y"									
# 32	1,524	18	9	6,213	1.0%	50.0%	2.41	2.11	2,35519
IF CONTM_WATR="N".and.CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y"									
# 33	1,064	433	17	6,266	2.0%	4.0%	2.41	2.30	39.96757
IF SMOKE_NOW="Y".and.OTHR_PAINT="Y".and.WOUNDED="Y" THEN FATIGUE="N".and.DJD/OSTEOARTHRITIS="N"									
# 34	23	6,543	22	1,202	96.0%	0.0%	2.40	2.14	40.12179
IF SMOKE_NOW="N".and.DIESEL_FUEL="N".and.CONTM_FOOD="Y" THEN ASTHMA="N".and.GERD="N"									
# 35	36	6,943	35	802	97.0%	1.0%	2.40	2.18	115.48438

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SMOKE_NOW="Y" .and. DIESEL_FUEL="N" .and. CHEM_ALARM="Y" THEN PAPULA ECZEMA="Y"									
# 36	45	345	7	7,363	16.0%	2.0%	2.39	2.07	7.66755
IF SERVICE="1" THEN MECHANICAL LOW BACK PAIN="Y" .and. MAJOR DEPRESSION="Y"									
# 37	6,274	18	17	1,471	0.0%	94.0%	2.39	2.10	21.63696
IF OTHER_PAINT="Y" .and. NONAF_FOOD="Y" .and. WOUNDED="N" THEN PERENNIAL ALLERGIC RHINITIS="Y" .and. SEVERE SLEEP APNEA="Y"									
# 38	2,200	18	11	5,539	1.0%	61.0%	2.38	1.91	0.85731
IF PYRIDOSTIG="N" .and. ACT_COMBAT="N" .and. RACE="C" THEN POST-TRAUMATIC STRESS DISORDER="Y"									
# 39	457	433	7	6,863	2.0%	2.0%	2.38	2.19	15.16222
IF SMOKE_NOW="Y" .and. DIESEL_FUEL="N" .and. NONAF_FOOD="Y" THEN PAPULA ECZEMA="Y"									
# 40	46	345	7	7,362	15.0%	2.0%	2.36	2.20	7.35993

Reference Number	Records Matching "IF" Matching "THEN" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.SEX="F"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.GERD="N"									
# 41	64	398	11	7,295	17.0%	3.0%	2.36	2.06	11.58398
IF PESTICIDES="Y".and.NONAF_FOOD="Y"									
THEN DJD/OSTEOARTHRITIS="Y".and.SEVERE SLEEP APNEA="Y"									
# 42	3,620	22	17	4,121	0.0%	77.0%	2.36	2.33	0.05543
IF SCUD_ATTAC="N".and.CHEM_ALARM="Y".and.PQ_AFTER="Y"									
THEN CHRONIC HEADACHES="Y".and.PATELLOFEMORAL PAIN SYNDROME="Y"									
# 43	208	53	5	7,490	2.0%	9.0%	2.35	2.12	5.49698
IF OTHER_PAINT="Y".and.NONAF_FOOD="Y"									
THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="Y"									
# 44	2,338	8	5	5,405	0.0%	63.0%	2.35	2.14	0.11993
IF OTHER_PAINT="Y"									
THEN SEVERE SLEEP APNEA="Y".and.MAJOR DEPRESSION="Y"									
# 45	3,393	8	6	4,351	0.0%	75.0%	2.35	2.70	0.03898

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.SEX="F"									
THEN POLY ARTHRALGIAS="N".and.MAJOR DEPRESSION="Y"									
# 46	64	172	5	7,515	8.0%	3.0%	2.34	2.25	5.61148
IF SERVICE="1"									
THEN PATELLOFEMORAL PAIN SYNDROME="Y".and.GERD="Y"									
# 47	6,274	34	32	1,470	1.0%	94.0%	2.33	2.79	42.47730
IF OTHR_PAINT="Y".and.RACE="R"									
THEN DEPRESSIVE DISORDER="N".and.MAJOR DEPRESSION="N"									
# 48	37	7,016	36	729	97.0%	1.0%	2.32	2.04	134.75497
IF OTHR_PAINT="N".and.NERVE_GAS="N"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.GERD="N"									
# 49	655	398	10	6,703	2.0%	3.0%	2.32	1.86	19.72983
IF OTHR_PAINT="N".and.CONTM_WATR="Y".and.RACE="H"									
THEN ASTHMA="N".and.INSOMNIA="N"									
# 50	7	6,988	5	756	71.0%	0.0%	2.31	2.58	26.62683

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SERVICE="4".and.NONAF_FOOD="Y".and.CASUALTIES="N" THEN DJD/OSTEOARTHRITIS="Y"									
# 51	67	397	11	7,293	16.0%	3.0%	2.31	2.27	10.66994
IF URANIUM="N".and.CHEM_ALARM="N" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.PERENNIAL ALLERGIC RHINITIS="N"									
# 52	738	414	12	6,606	2.0%	3.0%	2.30	2.18	23.32665
IF CONTM_WATR="Y".and.ACT_COMBAT="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.HYPERTENSION="Y"									
# 53	737	18	5	6,996	1.0%	28.0%	2.30	2.45	2.66555
IF OTHR_PAINT="Y".and.MALARIA="Y".and.WOUNDED="Y" THEN MEMORY LOSS="N".and.IRRITABLE BOWEL SYNDROME="Y"									
# 54	28	438	5	7,285	18.0%	1.0%	2.30	1.94	3.86459
IF OTHR_SOLVE="N".and.CHEM_ALARM="N".and.SEX="M" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.INSOMNIA="N"									
# 55	370	423	6	6,959	2.0%	1.0%	2.29	2.26	10.92871

Reference Number	Records Matching "IF" Matching "THEN" Statement		Records Matching Hypothesis		Records Not Matching Hypothesis		Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square									
IF CONTM_WATR="N".and.CASUALTIES="N"																				
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.MAJOR DEPRESSION="N"																				
# 56	1,064	389	17	6,310	2.0%	4.0%	2.29	2.28	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>17</td><td>1047</td></tr><tr><td>LHS false</td><td>372</td><td>6310</td></tr></table>		LHS true	RHS true	RHS false		17	1047	LHS false	372	6310	1064 6682
LHS true	RHS true	RHS false																		
	17	1047																		
LHS false	372	6310																		
IF SMOKE_PAST="N".and.NONAF_WATR="Y"																				
THEN DEPRESSIVE DISORDER="Y".and.PERENNIAL ALLERGIC RHINITIS="Y"																				
# 57	1,251	17	7	6,485	1.0%	41.0%	2.29	1.87	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>7</td><td>1244</td></tr><tr><td>LHS false</td><td>10</td><td>6485</td></tr></table>		LHS true	RHS true	RHS false		7	1244	LHS false	10	6485	1251 6495
LHS true	RHS true	RHS false																		
	7	1244																		
LHS false	10	6485																		
IF OTHR_SOLVE="Y".and.NONAF_FOOD="N".and.WOUNDED="Y"																				
THEN POLY ARTHRALGIAS="N".and.ASTHMA="N"																				
# 58	20	6,734	13	1,005	65.0%	0.0%	2.28	2.06	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>13</td><td>7</td></tr><tr><td>LHS false</td><td>6721</td><td>1005</td></tr></table>		LHS true	RHS true	RHS false		13	7	LHS false	6721	1005	20 7726
LHS true	RHS true	RHS false																		
	13	7																		
LHS false	6721	1005																		
IF URANIUM="Y".and.CONTM_FOOD="N".and.ANTHRAX="N"																				
THEN HEADACHES, MIGRAINE="N".and.IRRITABLE BOWEL SYNDROME="N"																				
# 59	27	6,811	26	934	96.0%	0.0%	2.28	2.54	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>26</td><td>1</td></tr><tr><td>LHS false</td><td>6785</td><td>934</td></tr></table>		LHS true	RHS true	RHS false		26	1	LHS false	6785	934	27 7719
LHS true	RHS true	RHS false																		
	26	1																		
LHS false	6785	934																		
# 59	27	6,811	26	934	96.0%	0.0%	2.28	2.54	<table><tr><td></td><td>6811</td><td>935</td></tr></table>			6811	935	70.10553						
	6811	935																		
IF OTHR_PAINT="Y".and.CASUALTIES="Y"																				
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.PERENNIAL ALLERGIC RHINITIS="Y"																				
# 60	2,168	19	11	5,570	1.0%	58.0%	2.27	2.16	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>11</td><td>2157</td></tr><tr><td>LHS false</td><td>8</td><td>5570</td></tr></table>		LHS true	RHS true	RHS false		11	2157	LHS false	8	5570	2168 5578
LHS true	RHS true	RHS false																		
	11	2157																		
LHS false	8	5570																		
# 60	2,168	19	11	5,570	1.0%	58.0%	2.27	2.16	<table><tr><td></td><td>19</td><td>7727</td></tr></table>			19	7727	0.72057						
	19	7727																		

Reference Number	Records Matching "IF" Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.SEX="F" THEN MAJOR DEPRESSION="Y"								
# 61	64	183	5	7,504	8.0%	3.0%	<div> <div>LHS true</div> <div>LHS false</div> </div> <div> <div>RHS true</div> <div>RHS false</div> </div>	<div>59</div> <div>7682</div>
							183	7563
							2.27	2.22
								4.94277
IF PQ_PRIOR="Y" THEN GERD="Y".and.HYPERTENSION="Y"								
							<div> <div>LHS true</div> <div>LHS false</div> </div> <div> <div>RHS true</div> <div>RHS false</div> </div>	<div>507</div> <div>7234</div>
							20	7214
							25	7721
# 62	512	25	5	7,214	1.0%	20.0%	2.27	1.87
								3.36231
IF SMOKE_NOW="N".and.OTHER_PAINT="Y".and.PQ_PRIOR="N" THEN CHRONIC FATIGUE="Y".and.HYPERTENSION="Y"								
							<div> <div>LHS true</div> <div>LHS false</div> </div> <div> <div>RHS true</div> <div>RHS false</div> </div>	<div>2020</div> <div>5721</div>
							4	5717
							9	7737
# 63	2,025	9	5	5,717	0.0%	56.0%	2.26	2.46
								0.25866
IF OTHER_PAINT="N".and.ANTHRAX="N".and.PQ_AFTER="Y" THEN MEMORY LOSS="N".and.INSOMNIA="N"								
							<div> <div>LHS true</div> <div>LHS false</div> </div> <div> <div>RHS true</div> <div>RHS false</div> </div>	<div>1</div> <div>7720</div>
							25	953
							6767	953
							6792	954
# 64	26	6,792	25	953	96.0%	0.0%	2.26	2.05
								65.54096
IF OTHER_PAINT="Y".and.MALARIA="Y".and.WOUNDED="Y" THEN IRRITABLE BOWEL SYNDROME="Y".and.PERENNIAL ALLERGIC RHINITIS="N"								
							<div> <div>LHS true</div> <div>LHS false</div> </div> <div> <div>RHS true</div> <div>RHS false</div> </div>	<div>23</div> <div>7718</div>
							5	7271
							447	7271
							452	7294
# 65	28	452	5	7,271	18.0%	1.0%	2.26	1.92
								3.58434

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SMOKE_PAST="Y".and.PASS_SMOKE="N".and.NERVE_GAS="N" THEN CHRONIC FATIGUE="N".and.MAJOR DEPRESSION="N"	58	7,298	57	447	98.0%	1.0%	2.26	2.03	368.59213
# 66	58	7,298	57	447	98.0%	1.0%	2.26	2.03	368.59213
IF SMOKE_NOW="Y".and.CASUALTIES="Y".and.RACE="H" THEN POST-TRAUMATIC STRESS DISORDER="Y"	66	433	11	7,258	17.0%	3.0%	2.24	1.81	9.05360
IF HEAT_SMOKE="N".and.PQ_PRIOR="Y".and.RACE="C" THEN CHRONIC FATIGUE="N".and.MAJOR DEPRESSION="N"	57	7,298	56	447	98.0%	1.0%	2.24	2.56	362.85744
IF SERVICE="3".and.CONTM_FOOD="N".and.ACT_COMBAT="N" THEN POLY ARTHRALGIAS="N"	39	7,101	38	644	97.0%	1.0%	2.24	2.43	166.18663
IF NERVE_GAS="Y" THEN FATIGUE="Y".and.POST-TRAUMATIC STRESS DISORDER="Y"	462	28	5	7,261	1.0%	18.0%	2.24	2.19	3.39642

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 71	IF WOUNDED="N" .and. CASUALTIES="N"								
	THEN DEPRESSIVE DISORDER="N" .and. POST-TRAUMATIC STRESS DISORDER="Y"								
	3,037	375	62	4,396	2.0%	17.0%	2.23	2.24	121.87190
# 72	IF OTHER_PAINT="Y" .and. MALARIA="Y" .and. WOUNDED="Y"								
	THEN IRRITABLE BOWEL SYNDROME="Y"								
	28	467	5	7,256	18.0%	1.0%	2.23	1.87	3.30605
# 73	IF SMOKE_NOW="Y" .and. DIESL_FUEL="Y" .and. WOUNDED="Y"								
	THEN POST-TRAUMATIC STRESS DISORDER="Y" .and. INSOMNIA="N"								
	43	423	7	7,287	16.0%	2.0%	2.23	1.90	5.34887
# 74	IF NERVE_GAS="Y" .and. CHEM_ALARM="Y" .and. PQ_PRIOR="Y"								
	THEN PATELLOFEMORAL PAIN SYNDROME="N" .and. SEVERE SLEEP APNEA="N"								
	33	7,009	32	736	97.0%	0.0%	2.22	2.20	119.05137
# 75	IF MICROWAVES="Y" .and. CONTM_WATR="Y"								
	THEN MEMORY LOSS="Y" .and. PERENNIAL ALLERGIC RHINITIS="Y"								
	696	20	5	7,035	1.0%	25.0%	2.22	1.94	2.45017

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SERVICE="3".and.OTHR_PAINT="N" THEN POLY ARTHRALGIAS="N"									
# 76	38	7,101	37	644	97.0%	1.0%	2.22	2.33	162.08870
IF DIESEL_FUEL="N".and.NERVE_GAS="Y" THEN POLY ARTHRALGIAS="N".and.SEVERE SLEEP APNEA="N"									
						LHS true LHS false	RHS true 7101	RHS false 1 645	38 7708
# 77	28	6,894	27	851	96.0%	0.0%	2.21	1.97	83.09590
IF CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.MEMORY LOSS="N"									
						LHS true LHS false	RHS true 420	RHS false 72 2998 348 4328 7326	3070 4676
# 78	3,070	420	72	4,328	2.0%	17.0%	2.21	2.29	137.80184
IF SMOKE_NOW="Y".and.BOTULISM="N".and.PQ_PRIOR="Y" THEN POLY ARTHRALGIAS="N".and.SEVERE SLEEP APNEA="N"									
						LHS true LHS false	RHS true 6894	RHS false 27 1 851 852	28 7718
# 79	28	6,894	27	851	96.0%	0.0%	2.21	1.96	83.09590
IF ANTHRAX="Y".and.WOUNDED="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.DJD/OSTEOARTHRITIS="N"									
						LHS true LHS false	RHS true 417	RHS false 12 64 405 7265 7329	76 7670
# 80	76	417	12	7,265	16.0%	3.0%	2.21	2.35	9.71855

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF CARC_PAINT="N".and.CONTM_WATR="N"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.CHRONIC HEADACHES="N"									
# 81	762	366	12	6,630	2.0%	3.0%	2.21	1.93	19.61709
IF OTHR_PAINT="Y".and.NONAF_FOOD="N".and.WOUNDED="Y"									
THEN POLY ARTHRALGIAS="N".and.ASTHMA="N"									
# 82	18	6,734	12	1,006	67.0%	0.0%	2.21	2.02	53.66585
IF SERVICE="2".and.NERVE_GAS="N"									
THEN PATELLOFEMORAL PAIN SYNDROME="Y"									
# 83	229	528	5	6,994	2.0%	1.0%	2.21	1.93	7.70813
IF NERVE_GAS="N".and.CHEM_ALARM="N"									
THEN POST-TRAUMATIC STRESS DISORDER="Y".and.PATELLOFEMORAL PAIN SYNDROME="N"									
# 84	751	402	13	6,606	2.0%	3.0%	2.21	2.16	21.39604
IF SERVICE="3"									
THEN INSOMNIA="Y"									
# 85	321	380	5	7,050	2.0%	1.0%	2.21	2.38	7.74551

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square		
# 86	IF NERVE_GAS="N".and.CONTM_FOOD="Y".and.CONTM_WATR="N" THEN CHRONIC FATIGUE="N".and.MAJOR DEPRESSION="N"				98.0%	1.0%	2.20	2.07	351.32407		
	55	7,298	54	447							
	<table><tr><td>LHS true</td><td>54</td><td>RHS true</td><td>1</td></tr><tr><td>LHS false</td><td>7244</td><td>RHS false</td><td>447</td></tr></table>		LHS true	54						RHS true	1
LHS true	54	RHS true	1								
LHS false	7244	RHS false	447								
# 87	IF CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.HEADACHES, MIGRAINE="N"				2.0%	17.0%	2.20	2.19	129.10230		
	3,070	396	68	4,348							
	<table><tr><td>LHS true</td><td>68</td><td>RHS true</td><td>3002</td></tr><tr><td>LHS false</td><td>328</td><td>RHS false</td><td>4348</td></tr></table>		LHS true	68						RHS true	3002
LHS true	68	RHS true	3002								
LHS false	328	RHS false	4348								
# 88	IF URANIUM="Y".and.MALARIA="Y".and.SCUD_ATTAC="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y"				16.0%	7.0%	2.20	1.84	23.22449		
	197	433	31	7,147							
	<table><tr><td>LHS true</td><td>31</td><td>RHS true</td><td>166</td></tr><tr><td>LHS false</td><td>402</td><td>RHS false</td><td>7147</td></tr></table>		LHS true	31						RHS true	166
LHS true	31	RHS true	166								
LHS false	402	RHS false	7147								
# 89	IF CASUALTIES="N" THEN FATIGUE="N".and.POST-TRAUMATIC STRESS DISORDER="Y"				2.0%	17.0%	2.20	2.25	131.56819		
	3,070	405	70	4,341							
	<table><tr><td>LHS true</td><td>70</td><td>RHS true</td><td>3000</td></tr><tr><td>LHS false</td><td>335</td><td>RHS false</td><td>4341</td></tr></table>		LHS true	70						RHS true	3000
LHS true	70	RHS true	3000								
LHS false	335	RHS false	4341								
# 90	IF SMOKE_NOW="Y".and.DIESL_FUEL="Y".and.WOUNDED="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y"				16.0%	2.0%	2.20	2.03	5.05514		
	43	433	7	7,277							
	<table><tr><td>LHS true</td><td>7</td><td>RHS true</td><td>36</td></tr><tr><td>LHS false</td><td>426</td><td>RHS false</td><td>7277</td></tr></table>		LHS true	7						RHS true	36
LHS true	7	RHS true	36								
LHS false	426	RHS false	7277								

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square					
# 91	IF URANIUM="Y".and.CONTM_WATR="Y".and.WOUNDED="Y" THEN MECHANICAL LOW BACK PAIN="Y"				23.0%	7,078	6	648	26 7720					
# 92	IF ANTHRAX="Y".and.WOUNDED="Y" THEN POST-TRAUMATIC STRESS DISORDER="Y".and.INSOMNIA="N"				16.0%	7,259	12	423	76 7670					
# 93	IF PQ_PRIOR="N".and.PQ_AFTER="Y" THEN MECHANICAL LOW BACK PAIN="N".and.SEVERE SLEEP APNEA="Y"				1.0%	6,936	5	192	623 7123					
# 94	IF CASUALTIES="N" THEN POST-TRAUMATIC STRESS DISORDER="Y"				2.0%	4,318	75	433	3070 4676					
# 95	IF SEX="F" THEN ASTHMA="Y".and.IRRITABLE BOWEL SYNDROME="Y"				1.0%	6,816	7	23	914 6832					

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square		
# 96	IF SMOKE_NOW="Y".and.DIESL_FUEL="N"				11	7,336	12.0%	3.0%	2.17	1.97	8.88582
	THEN ASTHMA="N".and.PAPULA_ECZEMA="Y"										
	89	332									
# 97	IF SERVICE="3"				5	7,064	2.0%	1.0%	2.17	2.32	7.16304
	THEN INSOMNIA="Y".and.HYPERTENSION="N"										
	321	366									
# 98	IF OTHER_PAINT="Y".and.ACT_COMBAT="Y".and.RACE="H"				11	7,254	16.0%	3.0%	2.16	1.84	7.97748
	THEN POST-TRAUMATIC STRESS DISORDER="Y"										
	70	433									
# 99	IF SMOKE_PAST="Y".and.WOUNDED="Y"				9	7,264	16.0%	2.0%	2.15	2.34	6.15691
	THEN POST-TRAUMATIC STRESS DISORDER="Y"										
	58	433									
# 100	IF BOTULISM="Y".and.CASUALTIES="N"				5	7,153	1.0%	20.0%	2.15	1.88	2.48759
	THEN DEPRESSIVE DISORDER="Y".and.INSOMNIA="Y"										
	573	25									



NPS Data Mining Initiative (DaMI)

09/06/96 Detailed Hypothesis Report: Exposure-to-symptom Study

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF MICROWAVES="Y".and.CONTM_FOOD="Y".and.RACE="H" THEN BLEED="Y".and.WEIGH="Y"								
# 1	38	127	5	7,586	13.0%	4.0%	RHS true 5 RHS false 33 122 7586 127 7619	38 7708
IF NERVE_GAS="Y".and.ACT_COMBAT="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.MUSCL="Y"								
# 2	22	308	6	7,422	27.0%	2.0%	RHS true 6 RHS false 16 302 7422 308 7438	22 7724
IF NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.WEIGH="Y"								
# 3	39	127	5	7,585	13.0%	4.0%	RHS true 5 RHS false 34 122 7585 127 7619	39 7707
IF HEAT_SMOKE="Y".and.MUSTRD_GAS="Y".and.NONAF_WATR="Y" THEN BLEED="Y".and.WEIGH="Y"								
# 4	56	127	7	7,570	13.0%	6.0%	RHS true 7 RHS false 49 120 7570 127 7619	56 7690
IF SMOKE_NOW="Y".and.MUSTRD_GAS="Y".and.NONAF_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"								
# 5	42	127	5	7,582	12.0%	4.0%	RHS true 5 RHS false 37 122 7582 127 7619	42 7704
# 5	42	127	5	7,582	12.0%	4.0%	3.13	2.80

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF OIL_SMOKE="Y".and.NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.MUSCL="Y"							RHS true RHS false		
							LHS true	7	
							LHS false	301	
								308	
# 6	30	308	7	7,415	23.0%	2.0%	3.01	2.56	17.41912
IF MICROWAVES="Y".and.MUSTRD_GAS="Y".and.NONAF_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	6	
							LHS false	121	
								127	
# 7	57	127	6	7,568	11.0%	5.0%	3.00	2.87	18.55428
IF SERVICE="1".and.NERVE_GAS="N".and.WOUNDED="Y" THEN DIFFI="N".and.MUSCL="N"							RHS true RHS false		
							LHS true	12	
							LHS false	4800	
								4812	
# 8	13	4,812	12	2,933	92.0%	0.0%	2.99	2.84	1.55608
IF OTHR_PAINT="Y".and.MUSTRD_GAS="Y".and.NONAF_WATR="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	5	
							LHS false	122	
								127	
# 9	49	127	5	7,575	10.0%	4.0%	2.95	2.43	14.29422
IF PASS_SMOKE="Y".and.NERVE_GAS="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	7	
							LHS false	120	
								127	
# 10	71	127	7	7,555	10.0%	6.0%	2.93	2.47	20.26519

Reference Number	Records Matching "If" Matching Statement	Records Matching "Then" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.SCUID_ATTAC="N".and.PQ_PRIOR="Y" THEN DIFFI="N".and.HEAD="N"									
# 11	8	3,923	7	3,822	88.0%	0.0%	2.92	2.42	0.00250
IF MICROWAVES="Y".and.MUSTRD_GAS="Y". THEN BLEED="Y".and.WEIGH="Y"									
# 12	72	127	7	7,554	10.0%	6.0%	2.91	2.39	19.87738
IF DIESEL_FUEL="Y".and.MUSTRD_GAS="Y".and.NONAF_WATR="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 13	72	127	7	7,554	10.0%	6.0%	2.91	2.50	19.87738
IF OIL_SMOKE="Y".and.MUSTRD_GAS="Y".and.NONAF_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 14	95	127	9	7,533	9.0%	7.0%	2.90	2.38	25.25639
IF NERVE_GAS="Y".and.MUSTRD_GAS="Y".and.SEX="M" THEN BLEED="Y".and.WEIGH="Y"									
# 15	73	127	7	7,553	10.0%	6.0%	2.90	2.39	19.49968

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF SMOKE_NOW="Y".and.MUSTRD_GAS="Y" THEN HAIRL="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	6	
							LHS false	134	
								140	7689
# 16	57	140	6	7,555	11.0%	4.0%	2.89	2.40	16.13193
IF MUSTRD_GAS="Y".and.CONTM_FOOD="Y".and.NONAF_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	5	
							LHS false	122	
								127	7694
# 17	52	127	5	7,572	10.0%	4.0%	2.89	2.37	13.16362
IF MICROWAVES="Y".and.CONTM_FOOD="Y".and.RACE="R" THEN DEPRE="N".and.HEAD="N"							RHS true RHS false		
							LHS true	7	
							LHS false	4038	
								4045	7738
# 18	8	4,045	7	3,700	88.0%	0.0%	2.86	3.08	0.02959
IF NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.MUSCL="Y"							RHS true RHS false		
							LHS true	8	
							LHS false	300	
								308	7707
# 19	39	308	8	7,407	21.0%	3.0%	2.85	2.43	17.10678
IF MUSTRD_GAS="Y".and.NONAF_WATR="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true RHS false		
							LHS true	7	
							LHS false	120	
								127	7670
# 20	76	127	7	7,550	9.0%	6.0%	2.85	2.42	18.42355

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF MUSTRD_GAS="Y".and.MALARIA="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true	RHS false	
							LHS true	5	49
							LHS false	122	7570
							127	7619	54 7692
# 21	54	127	5	7,570	9.0%	4.0%	2.85	2.60	12.47655
IF CARC_PAINT="Y".and.NERVE_GAS="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true	RHS false	
							LHS true	5	49
							LHS false	122	7570
							127	7619	54 7692
# 22	54	127	5	7,570	9.0%	4.0%	2.85	2.41	12.47655
IF PESTICIDES="Y".and.MUSTRD_GAS="Y".and.ANTHRAX="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true	RHS false	
							LHS true	7	70
							LHS false	120	7549
							127	7619	77 7689
# 23	77	127	7	7,549	9.0%	6.0%	2.84	2.55	18.08271
IF SERVICE="3".and.MUSTRD_GAS="Y" THEN DIFFI="Y"							RHS true	RHS false	
							LHS true	7	3
							LHS false	2143	5593
							2150	5596	10 7736
# 24	10	2,150	7	5,593	70.0%	0.0%	2.81	2.34	0.92505
IF SERVICE="3".and.DIESEL_FUEL="Y".and.MUSTRD_GAS="Y" THEN DIFFI="Y"							RHS true	RHS false	
							LHS true	7	3
							LHS false	2143	5593
							2150	5596	10 7736
# 25	10	2,150	7	5,593	70.0%	0.0%	2.81	2.34	0.92505

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.CHEM_ALARM="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.RASH="N"									
# 26	33	332	7	7,388	21.0%	2.0%	2.81	2.30	13.58639
IF NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.RASH="N"									
# 27	39	332	8	7,383	21.0%	2.0%	2.77	2.24	15.13160
IF NERVE_GAS="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 28	82	127	7	7,544	9.0%	6.0%	2.77	2.41	16.49891
IF SERVICE="5".and.PESTICIDES="Y" THEN DEPRE="N".and.MUSCL="N"									
# 29	11	4,948	10	2,797	91.0%	0.0%	2.73	2.32	1.74787
IF SMOKE_NOW="Y".and.MUSTRD_GAS="Y".and.BOTULISM="Y" THEN WEIGH="Y"									
# 30	24	538	7	7,191	29.0%	1.0%	2.72	2.18	9.26829

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 31	IF OIL_SMOKE="Y".and.MUSTRD_GAS="Y".and.ANTHRAX="Y" THEN BLEED="Y".and.WEIGH="Y"				8.0%	LHS true LHS false	RHS true RHS false	7 80 120 7539 127 7619	87 7659
					6.0%			2.70 2.45	15.09237
	IF MUSTRD_GAS="Y".and.NONAF_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"					LHS true LHS false	RHS true RHS false	9 105 118 7514 127 7619	114 7632
# 32					8.0%			2.70 2.36	19.32748
	IF MUSTRD_GAS="Y".and.ANTHRAX="Y" THEN BLEED="Y".and.WEIGH="Y"					LHS true LHS false	RHS true RHS false	8 92 119 7527 127 7619	100 7646
					6.0%			2.70 2.45	17.36376
# 33	IF PYRIDOSTIG="N".and.NONAF_WATR="N".and.RACE="M" THEN FATIG="N".and.MUSCL="N"				8.0%	LHS true LHS false	RHS true RHS false	5 1 3740 4000 3745 4001	6 7740
					0.0%			2.68 2.31	0.01201
	IF SERVICE="X".and.PYRIDOSTIG="N".and.CASUALTIES="N" THEN FATIG="N".and.ABDOM="N"					LHS true LHS false	RHS true RHS false	5 1 3778 3962 3783 3963	6 7740
# 34					83.0%			2.66 3.00	0.00594
					0.0%				
# 35					83.0%			2.66 3.00	0.00594
					0.0%				

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF PASS_SMOKE="N".and.URANIUM="Y".and.CONTM_WATR="Y" THEN DIARR="Y".and.WEIGH="Y"							RHS true	RHS false	
							LHS true	5	30
							LHS false	237	7474
# 36	35	242	5	7,474	14.0%	2.0%	2.66	2.19	8.48302
IF URANIUM="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"							RHS true	RHS false	
							LHS true	5	60
							LHS false	122	7559
# 37	65	127	5	7,559	8.0%	4.0%	2.64	2.49	9.43330
IF MUSTRD_GAS="Y".and.CONTM_FOOD="N".and.MALARIA="N" THEN DIARR="N".and.HEAD="N"							RHS true	RHS false	
							LHS true	6	1
							LHS false	4170	3569
# 38	7	4,176	6	3,569	86.0%	0.0%	2.64	2.18	0.07994
IF NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.JOINT="Y"							RHS true	RHS false	
							LHS true	10	29
							LHS false	490	7217
# 39	39	500	10	7,217	26.0%	2.0%	2.63	2.33	13.20972
IF MUSTRD_GAS="N".and.NONAF_WATR="Y".and.RACE="R" THEN RASH="N".and.SLEEP="N"							RHS true	RHS false	
							LHS true	5	1
							LHS false	3854	3886
# 40	6	3,859	5	3,886	83.0%	0.0%	2.62	2.83	0.00014

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square												
# 41	IF MUSTRD_GAS="Y".and.CONTM_FOOD="Y".and.CONTM_WATR="Y" THEN HAIRL="Y".and.WEIGH="Y"				8.0%	7,550	4.0%	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>5</td><td>56</td></tr><tr><td>LHS false</td><td>135</td><td>7550</td></tr><tr><td></td><td>140</td><td>7606</td></tr></table>	LHS true	RHS true	RHS false		5	56	LHS false	135	7550		140	7606	61 7685
	LHS true	RHS true	RHS false																		
		5	56																		
	LHS false	135	7550																		
	140	7606																			
# 42	IF MUSTRD_GAS="Y".and.WOUNDED="Y" THEN ABDOM="Y".and.HEAD="Y"				38.0%	6,867	1.0%	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>5</td><td>8</td></tr><tr><td>LHS false</td><td>866</td><td>6867</td></tr><tr><td></td><td>871</td><td>6875</td></tr></table>	LHS true	RHS true	RHS false		5	8	LHS false	866	6867		871	6875	13 7733
	LHS true	RHS true	RHS false																		
		5	8																		
	LHS false	866	6867																		
	871	6875																			
# 43	IF URANIUM="N".and.NERVE_GAS="Y".and.MUSTRD_GAS="Y" THEN SLEEP="Y"				72.0%	5,057	0.0%	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>13</td><td>5</td></tr><tr><td>LHS false</td><td>2671</td><td>5057</td></tr><tr><td></td><td>2684</td><td>5062</td></tr></table>	LHS true	RHS true	RHS false		13	5	LHS false	2671	5057		2684	5062	18 7728
	LHS true	RHS true	RHS false																		
		13	5																		
	LHS false	2671	5057																		
	2684	5062																			
# 44	IF MUSTRD_GAS="Y".and.ACT_COMBAT="Y".and.WOUNDED="Y" THEN MUSCL="Y"				58.0%	5,997	0.0%	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>7</td><td>5</td></tr><tr><td>LHS false</td><td>1737</td><td>5997</td></tr><tr><td></td><td>1744</td><td>6002</td></tr></table>	LHS true	RHS true	RHS false		7	5	LHS false	1737	5997		1744	6002	12 7734
	LHS true	RHS true	RHS false																		
		7	5																		
	LHS false	1737	5997																		
	1744	6002																			
# 45	IF URANIUM="Y".and.WOUNDED="Y" THEN RASH="Y".and.WEIGH="Y"				14.0%	7,447	2.0%	<table><tr><td>LHS true</td><td>RHS true</td><td>RHS false</td></tr><tr><td></td><td>6</td><td>36</td></tr><tr><td>LHS false</td><td>257</td><td>7447</td></tr><tr><td></td><td>263</td><td>7483</td></tr></table>	LHS true	RHS true	RHS false		6	36	LHS false	257	7447		263	7483	42 7704
	LHS true	RHS true	RHS false																		
		6	36																		
	LHS false	257	7447																		
	263	7483																			

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 51	IF MICROWAVES="Y".and.NERVE_GAS="Y".and.ANTHRAX="Y" THEN BLEED="Y".and.WEIGH="Y"								
	147	127	10	7,482	7.0%	8.0%	2.54	2.25	16.99236
# 52	IF PYRIDOSTIG="N".and.NONAF_WATR="N".and.RACE="M" THEN FATIG="N".and.WEIGH="N"								
# 53	IF MUSTRD_GAS="Y".and.CASUALTIES="Y".and.CHEM_ALARM="Y" THEN BLEED="Y".and.WEIGH="Y"								
	102	127	7	7,524	7.0%	6.0%	2.53	2.07	11.69246
# 54	IF SMOKE_NOW="Y".and.MICROWAVES="Y".and.MALARIA="Y" THEN MUSCL="Y".and.WEIGH="Y"								
	197	270	26	7,305	13.0%	10.0%	2.52	2.03	36.68523
# 55	IF SERVICE="4".and.OIL_SMOKE="N".and.WOUNDED="N" THEN FATIG="N".and.DIARR="N"								
	16	3,779	13	3,964	81.0%	0.0%	2.52	2.26	0.05839

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 56	IF MICROWAVES="Y".and.NERVE_GAS="Y".and.BOTULISM="Y" THEN BLEED="Y".and.WEIGH="Y"				103	127	7	RHS true 7 RHS false 96	103
							120	7523	7643
							127	7619	
							6.0%	2.52	2.15
# 57	IF PYRIDOSTIG="N".and.CONTM_FOOD="Y".and.BOTULISM="Y" THEN BLEED="N".and.WEIGH="Y"				25	411	5	RHS true 5 RHS false 20	25
							406	7315	7721
							411	7335	
							1.0%	2.51	2.02
# 58	IF PASS_SMOKE="Y".and.MUSTRD_GAS="Y".and.CONTM_WATR="Y" THEN SHORT="Y".and.WEIGH="Y"				67	169	6	RHS true 6 RHS false 61	67
							163	7516	7679
							169	7577	
							4.0%	2.51	2.27
# 59	IF OIL_SMOKE="Y".and.NERVE_GAS="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.RASH="N"				30	332	5	RHS true 5 RHS false 25	30
							327	7389	7716
							332	7414	
							2.0%	2.51	2.37
# 60	IF MUSTRD_GAS="Y".and.CHEM_ALARM="Y" THEN BLEED="Y".and.WEIGH="Y"				135	127	9	RHS true 9 RHS false 126	135
							118	7493	7611
							127	7619	
							7.0%	2.51	2.11

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Verification	Chi-square
IF PASS_SMOKE="Y".and.MUSTRD_GAS="Y".and.PQ_AFTER="Y" THEN BLEED="Y"									
# 61	17	665	5	7,069	29.0%	1.0%	2.50	2.11	3.90303
IF CARC_PAINT="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 62	105	127	7	7,521	7.0%	6.0%	2.50	2.24	11.12975
IF PASS_SMOKE="N".and.PYRIDOSTIG="N".and.ACT_COMBAT="Y" THEN FATIG="N".and.HEAD="N"									
# 63	12	3,134	9	4,609	75.0%	0.0%	2.49	2.77	0.00000
IF OTHR_PAINT="Y".and.WOUNDED="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 64	75	127	5	7,549	7.0%	4.0%	2.49	2.41	7.43650
IF OIL_SMOKE="Y".and.CARC_PAINT="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 65	90	127	6	7,535	7.0%	5.0%	2.49	2.15	9.29857

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF NERVE_GAS="Y".and.MALARIA="Y" THEN HAIRL="Y".and.MUSCL="Y"								
# 66	111	359	19	7,295	17.0%	5.0%	<div>RHS true</div> <div>RHS false</div> <div>19</div> <div>340</div> <div>7295</div> <div>7387</div>	<div>111</div> <div>7635</div>
IF PASS_SMOKE="Y".and.MUSTRD_GAS="Y". THEN BLEED="Y".and.WEIGH="Y"								
# 67	155	127	10	7,474	6.0%	8.0%	<div>RHS true</div> <div>RHS false</div> <div>10</div> <div>117</div> <div>145</div> <div>7474</div> <div>7619</div>	<div>155</div> <div>7591</div>
IF NERVE_GAS="N".and.CONTM_FOOD="N" THEN DIFFI="Y".and.WEIGH="Y"								
# 68	896	302	9	6,557	1.0%	3.0%	<div>RHS true</div> <div>RHS false</div> <div>9</div> <div>293</div> <div>887</div> <div>6557</div> <div>7444</div>	<div>896</div> <div>6850</div>
IF SMOKE_PAST="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"								
# 69	76	127	5	7,548	7.0%	4.0%	<div>RHS true</div> <div>RHS false</div> <div>5</div> <div>122</div> <div>71</div> <div>7548</div> <div>7619</div>	<div>76</div> <div>7670</div>
IF SERVICE="X".and.HEAT_SMOKE="N" THEN HAIRL="N".and.MUSCL="N"								
# 70	11	5,402	10	2,343	91.0%	0.0%	<div>RHS true</div> <div>RHS false</div> <div>10</div> <div>5392</div> <div>1</div> <div>2343</div> <div>2344</div>	<div>11</div> <div>7735</div>
						2.47	2.69	3.86408

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 71	IF MUSTRD_GAS="Y".and.CONTM_FOOD="Y".and.SEX="M" THEN HAIRL="Y".and.WEIGH="Y"				7.0%	4.0%	2.47	2.43	7.22349
	69	140	5	7,542					
# 72	IF SMOKE_PAST="Y".and.MUSTRD_GAS="Y".and.SEX="M" THEN HAIRL="Y".and.WEIGH="Y"				7.0%	4.0%	2.46	2.45	7.04285
	70	140	5	7,541					
# 73	IF SMOKE_PAST="N".and.CHEM_ALARM="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.WEIGH="Y"				6.0%	8.0%	2.46	2.16	14.99895
	158	127	10	7,471					
# 74	IF PYRIDOSTIG="N".and.NONAF_WATR="N".and.RACE="M" THEN FATIG="N"				83.0%	0.0%	2.46	2.52	0.06195
	6	4,163	5	3,582					
# 75	IF MUSTRD_GAS="Y".and.CONTM_FOOD="Y" THEN BLEED="Y".and.WEIGH="Y"				6.0%	4.0%	2.46	1.97	7.10045
	77	127	5	7,547					

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF WOUNDED="Y".and.CHEM_ALARM="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 76	94	127	6	7,531	6.0%	5.0%	2.45	1.99	8.61631
IF OIL_SMOKE="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 77	143	127	9	7,485	6.0%	7.0%	2.45	2.04	13.28653
IF MUSTRD_GAS="Y".and.SCUDD_ATTAC="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 78	110	127	7	7,516	6.0%	6.0%	2.45	2.30	10.26171
IF CONTM_FOOD="Y".and.PQ_PRIOR="Y" THEN DIARR="Y".and.WEIGH="Y"									
# 79	94	242	11	7,421	12.0%	5.0%	2.45	2.06	15.35964
IF MUSTRD_GAS="Y".and.NONAF_FOOD="Y".and.MALARIA="Y" THEN ABDOM="Y".and.DEPRE="Y"									
# 80	34	614	9	7,107	26.0%	1.0%	2.44	2.53	8.07940

Reference Number	Records Matching "IF" Matching Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
# 81	IF MUSTRD_GAS="N".and.CONTM_FOOD="N".and.NONAF_FOOD="Y" THEN DIFFI="Y".and.WEIGH="Y"				771	302	8	RHS true 8	771
					8	6,681	294	RHS false 6681	6975
							302	7444	
# 82	IF URANIUM="Y".and.NONAF_WATR="Y".and.MALARIA="Y" THEN BLEED="Y".and.WEIGH="Y"				771	302	7	RHS true 7	113
					8	6,681	120	RHS false 7513	7633
							127	7619	
# 83	IF CARC_PAINT="N".and.CONTM_WATR="N" THEN DIFFI="Y".and.WEIGH="Y"				113	127	8	RHS true 8	762
					7	7,513	294	RHS false 754	6984
							302	7444	
# 84	IF SERVICE="X".and.OTHER_SOLVE="Y" THEN DIFFI="Y".and.SHORT="N"				762	302	6	RHS true 6	12
					8	6,690	1513	RHS false 6221	7734
							1519	6227	
# 85	IF OTHER_PAINT="N".and.CONTM_FOOD="N".and.RACE="N" THEN DEPRE="Y".and.SLEEP="Y"				12	1,519	6	RHS true 6	1,35688
					6	6,221	1259	RHS false 6211	276
							1272	6474	7470
# 86					276	1,272	13	RHS true 13	30,60700
					13	6,211	2,41	RHS false 2,41	1,97
							2,41	1,97	30,60700

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square
IF MUSTRD_GAS="Y".and.NONAF_WATR="Y".and.ACT_COMBAT="Y" THEN BLEED="Y".and.MUSCL="Y"									
# 86	42	308	6	7,402	14.0%	2.0%	2.41	2.36	6.83033
IF OTHR_SOLVE="Y".and.WOUNDED="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 87	81	127	5	7,543	6.0%	4.0%	2.40	2.32	6.47991
IF MUSTRD_GAS="Y".and.WOUNDED="Y" THEN MUSCL="Y"									
# 88	13	1,744	7	5,996	54.0%	0.0%	2.39	2.11	1.13285
IF URANIUM="N".and.CONTM_WATR="N" THEN BLEED="Y".and.WEIGH="Y"									
# 89	1,074	127	5	6,550	0.0%	4.0%	2.38	1.93	10.64825
IF SMOKE_NOW="N".and.NERVE_GAS="Y".and.MALARIA="Y" THEN ABDOM="Y".and.HAIRL="Y"									
# 90	76	291	10	7,389	13.0%	3.0%	2.38	2.28	12.00188

Reference Number	Records Matching "If" Statement	Records Matching "Then" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Association Verification	Chi-square				
# 91	IF MUSTRD_GAS="Y".and.WOUNDED="Y" THEN MUSCL="Y".and.SLEEP="Y"				38.0%	0.0%	2.38	2.46	1.96398				
	13	1,055	5	6,683									
	RHS true									RHS false			
	5									8			
# 92	IF MUSTRD_GAS="Y".and.ACT_COMBAT="Y".and.SCUD_ATTAC="Y" THEN BLEED="Y".and.MUSCL="Y"				14.0%	3.0%	2.38	2.20	10.45683				
	65	308	9	7,382									
	RHS true									RHS false			
	9									56			
# 93	IF MICROWAVES="Y".and.MUSTRD_GAS="Y" THEN BLEED="Y".and.MUSCL="Y"				14.0%	3.0%	2.38	2.73	11.84089				
	72	308	10	62									
	RHS true									RHS false			
	10									62			
# 94	IF OIL_SMOKE="Y".and.NERVE_GAS="N".and.WOUNDED="Y" THEN HAIRL="N".and.RASH="N"				88.0%	0.0%	2.37	2.36	1.26223				
	8	4,953	7	2,792									
	RHS true									RHS false			
	7									1			
# 95	IF MICROWAVES="Y".and.WOUNDED="Y" THEN SHORT="Y".and.WEIGH="Y"				8.0%	3.0%	2.37	2.13	5.98519				
	63	169	5	7,519									
	RHS true									RHS false			
	5									58			

Reference Number	Records Matching "IF" Statement	Records Matching "THEN" Statement	Records Matching Hypothesis	Records Not Matching Hypothesis	Forward Confidence Factor	Reverse Confidence Factor	Complex Association Factor	Complex Verification	Chi-square
IF CONTM_WATR="Y".and.WOUNDED="Y" THEN DIFFI="Y".and.WEIGH="Y"									
# 96	59	302	8	7,393	14.0%	3.0%	2.37	2.20	9.13784
IF NERVE_GAS="Y".and.PQ_AFTER="Y".and.RACE="C" THEN ABDOM="Y"									
# 97	27	1,331	12	6,400	44.0%	1.0%	2.36	1.98	4.08247
IF ANTHRAX="Y".and.SCUDD_ATTAC="Y".and.PQ_PRIOR="Y" THEN BLEED="Y".and.WEIGH="Y"									
# 98	155	127	9	7,473	6.0%	7.0%	2.36	1.94	11.46153
IF NERVE_GAS="Y".and.WOUNDED="Y" THEN ABDOM="Y".and.DIFFI="Y"									
# 99	22	693	6	7,037	27.0%	1.0%	2.35	2.35	3.87528
IF NERVE_GAS="N".and.CONTM_WATR="N".and.ANTHRAX="N" THEN ABDOM="Y".and.MUSCL="Y"									
# 100	261	630	6	6,861	2.0%	1.0%	2.35	1.99	12.20504

[THIS PAGE INTENTIONALLY LEFT BLANK]

LIST OF REFERENCES

- Antonoff, Michael, "Genetic algorithms: software by natural selection," *Popular Science*, vol. 239, no. 4, pp. 101-105, October 1991.
- CCEP, "CCEP Report on 18,598 Participants," *Comprehensive Clinical Evaluation Program for Persian Gulf War Veterans*, April 2, 1996.
- CCEP, "Summary of interview with Director, Deployment Surveillance Team and CCEP Epidemiologists," September 11, 1996.
- Davis, Lawrence, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, 1991.
- Dixon, Wilfrid J. and Frank J. Massey, Jr., *Introduction to Statistical Analysis*, Magraw-Hill, 1969.
- Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, 1996.
- Fletcher, Robert H., *Clinical Epidemiology—The Essentials*, Williams & Wilkins, 1982.
- Goldberg, David E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989.
- Holland, John, H., *Adaptation in Natural and Artificial Systems*, The MIT Press, 1975.
- Joseph, SC. *Memorandum to Deputy Secretary of Defense. Subject: Persian Gulf Syndrome - Health Issues of Military Personnel*, May 11, 1994.
- Koza, John R., *Genetic Programming, On the Programming of Computers by Means of Natural Selection*, The MIT Press, 1990.
- Walpole, Ronald E. and Raymond H. Meyers, *Probability and Statistics for Engineering Science*, Macmillan, 1988.

[THIS PAGE INTENTIONALLY LEFT BLANK]

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center2
8725 John J. Kingham Road., Ste 0944
Ft. Belvoir, Virginia 22060-6218

2. Dudley Knox Library2
Naval Postgraduate School
411 Dyer Rd.
Monterey, California 93943-5101

3. Assistant Secretary of Defense (Health Affairs)2
1200 Defense Pentagon
Washington, DC 20301-1200

4. Deployment Surveillance Team10
5205 Leesburgh Pike, Suite 1135
Falls Church, VA 22041

5. Commanding Officer, Naval Medical Information Management Center1
8901 Wisconsin Avenue
Building 27
Bethesda, MD 20889-5605

6. Dr. Jim Emery1
Associate Provost for Computer and Information Services
Root Hall, Room 265
Naval Postgraduate School
Monterey, CA 93943

7. Dr. Hemant Bhargava, Code SM/BH5
Department of Systems Management
Naval Postgraduate School
Monterey, California 93943

8. Dr. Donald P. Gaver, Code OR/G2
Department of Operations Research
Naval Postgraduate School
Monterey, California 93943

9. Dr. Steve R. Lamar, Code 01.....1
Executive Director, Institute for Defense Education and Analysis
Root 217
Naval Postgraduate School
Monterey, CA 93943
10. LT David L. Jacobson.....3
1317 Neck Road
Burlington, NJ 08016
11. LT Debra A. Lankhorst.....1
P.O. Box 33
Chest Springs, PA 16624